Effects of recombination rate and mating system on genome evolution and diversity in *Arabidopsis*

Stephen Isaac Wright

Thesis presented for the degree of Doctor of Philosophy at the

University of Edinburgh

2003

Abstract

High levels of inbreeding are expected to cause a strong reduction in levels of genetic variability, effective recombination rates and in adaptation compared with related outcrossing populations. The evolution of mating systems can thus have profound effects on the evolution of genome structure and diversity. In this thesis I test these predictions, using the plant genus *Arabidopsis* as a model system. I examine patterns of genome organisation, DNA sequence polymorphism and divergence in the highly self-fertilizing *Arabidopsis thaliana*, and compare them to those of its self-incompatible, outcrossing relative *Arabidopsis lyrata*. From comparisons of rates of substitution, there is no evidence for a higher rate of amino acid substitution in *A. thaliana*, suggesting that slightly deleterious amino acid mutations may not be the primary source of protein evolution in these species. In contrast with results from published data, analysis including polymorphism data also reveals no difference in the ratio of nonsynonymous to synonymous polymorphism between species, suggesting that there may not be a general elevation of amino acid polymorphism in *A. thaliana*. Comparisons of intron length reveal evidence for consistently smaller introns in *A. thaliana*, perhaps reflecting the action of directional selection on noncoding DNA length in an annual plant.

Analysis of codon usage bias at orthologous loci shows evidence for consistently higher GC content at third codon positions in *A. lyrata*. Comparisons of base composition in introns and polymorphism patterns for preferred and unpreferred synonymous mutations show no evidence for a shift in mutation pattern or rates of biased gene conversion between species, suggesting that the difference in codon bias might reflect a relaxation of natural selection in *A. thaliana*. However,

comparisons of codon usage between species using a measure of codon bias based on relative abundance of tRNA genes reveals no significant difference between species, and there is no evidence for a difference in the relative rates of preferred and unpreferred substitutions. As there is a good correlation between the frequency of preferred codons defined by tRNA abundance and levels of gene expression, these results suggest a neutral explanation for the difference in GC content.

Comparisons of multilocus neutral variability between *A. thaliana* and *A. lyrata* show the expected decrease in average within-population diversity in *A. thaliana*, but this is complicated by strong geographic structuring of variability in *A. lyrata*, probably associated with recent demographic history. Consistent with an influence of demographic history, analysis of intralocus linkage disequilibrium suggests a strong deficiency of the effective rate of recombination in *A. lyrata*. In contrast, *A. thaliana* shows approximately the amount of linkage disequilibrium expected in a highly self-fertilising species. These results suggest a possible role for population admixture in northern, postglacial populations of *A. lyrata.*

A whole-genome analysis of transposable elements in *A. thaliana* indicates that recombination rate heterogeneity does not play an important role in driving their distribution in this species, and that gene density is the primary determinant of TE abundance in the genome. Combined with evidence from other complete eukaryotic genomes, this pattern is consistent with a role of inbreeding in reducing effective rates of recombination genome-wide, and thereby reducing the effect of recombination rate heterogeneity on genome structure.

Polymorphism analysis of 18 loci located in regions of contrasting recombination rate in an Icelandic population of *Arabidopsis lyrata* reveals no

evidence for the expected positive correlation between recombination rate and nucleotide diversity. A maximum likelihood analysis of polymorphism and divergence for these data shows evidence for significantly elevated diversity compared with divergence at a centromeric locus, suggesting the possibility of balancing selection in this region. Alternatively, recent demographic history may have contributed to an uncoupling of the expected relationship between recombination and variability, and an inflation of the variance in diversity across loci. The results of this thesis provide some evidence for the evolution of genome structure between related Arabidopsis species, but no strong evidence for differences in the efficacy of natural selection. These results emphasize the importance of understanding the influence of population history on the action of natural selection at the molecular level.

Notes on style and authorship

All of the results chapters (Chapters 2-6) have been, or will be, prepared for submission into peer-reviewed journals, with the contributions of coauthors outlined below. The contents of Chapter 2 have been published in *Molecular Biology and Evolution* (Wright, S.I., Lauga, B. and D. Charlesworth. 2002. *Mol. Biol. Evol.* **9**(9):1407-1420.). B. Lauga contributed sequence data from *A. lyrata* and Brassica species for this publication, and D. Charlesworth played an important role in the design of the project, a supervisory role during data collection, and an editorial role during the writing up. I contributed to the design of the project, the collection of sequence data from *A. lyrata* and Capsella species, and all of the analyses and writing. The new sequence data in Chapter 3 were generated by myself, K. Yau and M. Loosely, and K. Yau also contributed some of the analysis of codon bias and substitution rates for several individual loci. The design of the study, the remainder of the analyses, and the writing of this chapter are all my own work. Chapter 4 was published in *Molecular Ecology* (Wright, S.I., Lauga, B. and D. Charlesworth. 2003. *Mol Ecol* **12**:1247-1269.). B. Lauga and I each contributed the collection of polymorphism data from three genes for this work, and I designed and conducted the analysis and writing. D. Charlesworth played an important supervisory role during the planning, collection and analysis of this data, and played an editorial role during the writing up. The contents of Chapter 5 were published in *Genome Research* (Wright, S.I., Agrawal, N. and T.E. Bureau. 2003. *Genome Res.* **13**:1897-1903.). I planned the study, designed the analysis and data extraction, participated collaboratively in checks and corrections of the data, and did all analyses and

writing. N. Agrawal wrote Pearl scripts for data extraction, and downloaded and extracted the data from databases using these scripts. T. Bureau played a supervisory and editorial role, and also contributed to the checks, corrections and revisions of the dataset for analysis.

The data collection, analysis and writing of Chapter 6 are my own work, with D. Charlesworth and I both contributing to the design of the project, and D. Charlesworth providing an editorial role during the writing up. B. Charlesworth proposed the initial idea for a maximum likelihood analysis of polymorphism and divergence developed in this chapter, and he played an important supervisory role during my design, development of the likelihood equations, implementation of the method, and investigation of the behaviour of the method using coalescent simulations.

Acknowledgements

I thank my supervisor, Deborah Charlesworth, for her tremendous generosity of time, ideas and support over the last three years. She provided a wealth of encouragement, suggestions and freedom, and created a fantastically stimulating academic environment for a PhD student.

I also thank other members of the ICAPB who contributed greatly to my scientific development during my time here. Peter Andolfatto, Doris Bachtrog, Brian Charlesworth, Toby Johnson, Gabriel Marais, and Peter Keightley in particular provided numerous suggestions and discussions that were very important during my thesis. Thanks also to the members of the Charlesworth labs for their friendship and support during my time here. My work has also benefited greatly from discussions with and comments from Jeff Wall, Molly Przeworski, Brandon Gaut, Tom Mitchell-Olds, and Dan Schoen.

I thank Magnus Nordborg, Honggang Zheng, Blake Meyers, Eli Stahl, Gabriel Marais, Outi Savolainen, Helmi Kuittinen and Esther Kamau, who all contributed unpublished data for some of the analyses in this thesis, and Herbert Hurka generously provided seed material from Capsella species.

Lastly I thank my wife Amy Wright, for her continuous support and encouragement.

Table of Contents

Chapter 1- Introduction

A review of the effects of population size, recombination and mating system on genome evolution and diversity

1.1 Introduction

The genomes of organisms vary in their rates and patterns of molecular evolution, including patterns of base composition (TARRIO *et al.* 2001), rates of protein evolution (KEIGHTLEY and EYRE-WALKER 2000), and rates of insertion and deletion in noncoding DNA (PETROV *et al.* 1996). Two primary explanations have been proposed for these patterns. First, mutation biases and rates may differ among species and genomic regions. Alternatively, there may be differences in the strength and/or efficacy of natural selection between genomic regions and species. The relative importance of the effects of mutation versus selection, and the strength and direction of any selective effects, are largely unclear (KLIMAN and HEY 2003; MARAIS *et al.* 2001; RODRIGUEZ-TRELLES *et al.* 1999; SMITH and EYRE-WALKER 2001).

Mutational explanations for patterns of genome evolution

If the mutations that are segregating in natural populations are predominantly neutral, the rates and patterns of molecular evolution will be driven primarily by mutational biases and rates, since neutral evolution is determined by the rates of each class of mutation. Mutational explanations have recently been proposed to explain patterns of codon usage bias across the genomes and among species of *C. elegans* and Drosophila (MARAIS *et al.* 2001; RODRIGUEZ-TRELLES *et al.* 1999), differences

among sequenced genomes in the frequency of simple sequence repeats (KATTI *et al.* 2001) and transposable element distributions in the genome of *C. elegans* (DURET *et al.* 2000). Striking differences in base composition across the human genome have also been interpreted as the effect of mutational bias (DURET and GALTIER 2000). Finally, differences among organisms in the sizes of non-coding regions have been suggested to reflect contrasts in the relative rates of insertions and deletions in different species (BENSASSON *et al.* 2001; PETROV and HARTL 1997), leading to the striking differences in eukaryotic genome size. Support for a mutation-based explanation for the above cases has been indirect, relying on comparisons with regions that are thought to be evolving neutrally. However, a recent attempt at a direct comparison of rates of insertions and deletions in *Arabidopsis thaliana* and tobacco (KIRIK *et al.* 2000) provided support for the possibility that species may evolve substantially different mutational biases.

Non-neutral explanations for genome evolution: the nearly neutral model and effects of hitchhiking

Alternatively, if many of the mutations segregating in natural populations are subject to weak natural selection, many of these differences in patterns of genome evolution may be explained by differences in the efficacy of natural selection among species and genomic regions. The rate of molecular evolution for mutations subject to selection is not solely determined by mutation rates, but also by the product of the selection coefficient (s) and the effective population size (N_e)(KIMURA 1983). If, as proposed by the nearly neutral theory, a large fraction of mutations are slightly deleterious (on the order of the reciprocal of the effective population size),

differences in either selection coefficients or effective population size can have substantial effects on the evolution of genomes (OHTA 1992). Reduced effective size increases the role of drift in determining the fate of mutations relative to selection, leading to higher frequencies of slightly deleterious mutations, and elevated rates of their fixation. Reduced efficacy of natural selection can thus occur in small populations.

In addition to changes in population size, differences in the efficacy of selection can be driven by contrasts in rates of crossing over, through the effects of natural selection on the fate of linked segregating sites (HILL and ROBERTSON 1966). Effects of natural selection on the diversity and fixation of linked loci (hereafter 'hitchhiking') have been divided into three processes, based on the direction and strength of selection; strong positive selection ('selective sweeps') (BARTON 2000; BRAVERMAN *et al.* 1995), strong negative selection ('background selection') (CHARLESWORTH *et al.* 1993), and the action of weak Hill-Robertson interference between many weakly-selected sites segregating simultaneously (MCVEAN and CHARLESWORTH 2000; TACHIDA 2000). The effects of background selection have been shown to approximate a reduction in effective population size, where the effective size is reduced to the proportion of the population that is free from deleterious mutations (CHARLESWORTH 1994; CHARLESWORTH *et al.* 1993). This fraction is a function of the deleterious mutation rate and the selection coefficient. Since the size of the genomic region will be a major determinant of the deleterious mutation rate, the effects will be stronger in regions of the genome and species with reduced crossing over. With recurrent positive selection and weak Hill-Robertson interference, effects on linked sites are more complex, but the magnitude of these

effects will also be strongly affected by the rate of crossing over (BRAVERMAN *et al.* 1995; MCVEAN and CHARLESWORTH 2000).

The predictions of hitchhiking are supported by the observation of a correlation between rates of crossing over and levels of neutral genetic variability in several species, including Drosophila (BEGUN and AQUADRO 1992; LANGLEY *et al.* 2000; LANGLEY *et al.* 1993), mouse (NACHMAN 1997), humans (NACHMAN *et al.* 1998), *C. elegans* (CUTTER and PAYSEUR 2003), tomato (STEPHAN and LANGLEY 1998), and maize (TENAILLON *et al.* 2001). However, the strength of the effect (BAUDRY *et al.* 2001), the consistency of the result with different measures of recombination rate (TENAILLON *et al.* 2002) and the relative importance of selective vs. neutral explanations for the pattern (HELLMANN *et al.* 2003; LERCHER and HURST 2002) have recently been called into question for several species, and the general importance of natural selection in structuring patterns of diversity across the genome remains unclear. Furthermore, even in the genus *Drosophila*, where an important effect of natural selection on reducing variability in regions of low recombination is well-accepted, the relative contribution of different types of selection are unresolved (ANDOLFATTO 2001). A central approach in the attempt to distinguish between hitchhiking models lies in their differential effects on the frequency spectrum of segregating sites (BRAVERMAN *et al.* 1995; CHARLESWORTH *et al.* 1995; FAY and WU 2000). In particular, selective sweeps are expected to generate a significant skew towards rare variants, and in some cases an excess of high frequency derived variants, relative to that expected under neutrality.

1.2 Evidence for a role of weak purifying selection in driving patterns of molecular evolution

Rates of protein evolution

The above models predict that, if there is a large class of weakly selected mutations, then population size and recombination rate can have important effects on molecular evolution. The relatively low levels of amino acid polymorphism in natural populations of Drosophila (MORIYAMA and POWELL 1996) suggest that many amino acid mutations may be under strong purifying selection, and are unlikely to rise to high frequency in all but the smallest populations. However, the magnitude and direction of selection on those replacement mutations that do segregate and fix in natural populations is strikingly unresolved. One method for assessing selection on replacement mutations is a comparison of the ratio of polymorphism to fixation for amino acid to silent substitutions, using the McDonald-Kreitman (MK) test (MCDONALD and KREITMAN 1991; SAWYER and HARTL 1992). The neutrality index (WEINREICH and RAND 2000) has been used as a summary of this comparison:

$$\text{N.I.} = \frac{\text{(no. replacement polymorphisms/no. replacement fixations)}}{\text{(no.synonymous polymorphisms/no.synonymous fixations)}}$$

Excesses of replacement polymorphism relative to fixation (N.I. >1) are consistent with models of weak purifying selection, since natural selection is more effective at preventing fixation than at allowing segregating polymorphism (TACHIDA 2000) of slightly deleterious mutations. No difference in the numerator compared with the denominator (N.I.=1) would suggest neutrality, whereas an excess of replacement fixation (N.I.<1) would reflect the action of positive selection. In Drosophila, the

mitochondrial genome shows a pattern consistent with weak purifying selection (WEINREICH and RAND 2000). In *Arabidopsis thaliana*, summaries of patterns of polymorphism from a small set of nuclear genes also suggest a common excess of polymorphic replacement mutations, as expected from the nearly neutral model (WEINREICH and RAND 2000). However, among nuclear genes in Drosophila, a large amount of variation is observed in the MK test among loci, with many genes showing no significant difference or an excess of fixed replacement substitution (MORIYAMA and POWELL 1996; WEINREICH and RAND 2000). These multilocus patterns have been used to estimate a high rate of adaptive protein evolution in Drosophila (BUSTAMANTE *et al.* 2002; FAY *et al.* 2002; SMITH and EYRE-WALKER 2002), and the predominance of slightly deleterious amino acid mutations in *Arabdiopsis thaliana* (BUSTAMANTE et al. 2002), although these interpretations are based on a relatively restricted number of loci, and may be influenced by the presence of unusual genes sequenced by researchers with a priori hypotheses of selection.

Several studies have found evidence that population size can influence the rate of protein evolution. These have involved a comparison of the ratio of nonsynonymous to synonymous substitution in different lineages, under the assumption that synonymous mutations are neutral. Ohta (1993b) found support for the nearly neutral model from the inverse correlation between generation time and per-generation rates of protein evolution in mammals. Since organisms with longer generation times tend to have lower population sizes, this was interpreted as evidence for reduced efficacy of selection in small populations. More recently, a much larger sample of genes has provided further evidence for a generation time

effect on rates of replacement substitution (KEIGHTLEY and EYRE-WALKER 2000). Evidence for elevated rates of replacement relative to silent substitution at the ADH locus in Hawaiian Drosophila populations, which are thought to have undergone population bottlenecks following colonization (DESALLE 1988; OHTA 1993a), and at two mitochondrial loci in island bird species compared to mainland relatives (JOHNSON and SEGER 2001), are also broadly consistent with the nearly neutral model. Finally, comparison of rates of replacement relative to silent substitutions in the closely related *Drosophila melanogaster* and *Drosophila simulans* indicated an acceleration of amino acid substitution in *D. melanogaster* (AKASHI 1996). As *D. melanogaster* appears to have lower levels of silent diversity (MORIYAMA and POWELL 1996), this contrast may reflect effects of small reductions in population size.

Effects of recombination rate on rates of protein evolution have been less characterized, partly because of the difficulty of identifying reasonable comparisons, as selection coefficients on different genes vary substantially, and species with contrasting rates of recombination are often highly diverged from each other (WELCH and MESELSON 2001). Nevertheless, differences in rates of gene evolution between free-living bacteria and the endosymbiotic bacterium Buchnera (WERNEGREEN and MORAN 1999), and between organelle and nuclear genes (LYNCH and BLANCHARD 1998), are consistent with the prediction that slightly deleterious replacement mutations accumulate in non-recombining genomes. While the number of genes in these endosymbiont genomes may be too few to allow the effects of background selection, selective sweeps and weak Hill-Robertson interference may still be effective. Observed reductions in silent polymorphism in the Buchnera

genome (FUNK *et al.* 2001) support the underlying assumption of reduced effective population size. However, the endosymbiotic life history might also cause a reduction in selection coefficients, making it difficult to distinguish between effects on selection coefficient vs. effective size. In contrast with these results, Betancourt and Presgraves (2002) showed a higher rate of amino acid substitution in regions of high recombination in *Drosophila melanogaster*, suggesting that a reduced efficacy of natural selection in regions of low recombination can also slow the rate of adaptive substitution. Although this study included a large number of male-specific genes, a class of genes generally thought to be subject to adaptive protein evolution in Drosophila, this result shows that the presence of both adaptive and deleterious amino acid mutations can complicate a general test for the role of hitchhiking on rates protein evolution.

Codon usage bias

In addition to a subset of replacement mutations, many synonymous mutations are potential candidates for the action of weak selection. Several aspects of the patterns of synonymous codon usage are consistent with a model of mutation-selection-drift balance (AKASHI 1997). First, the level of codon usage bias has been shown to correlate with gene expression in *Drosophila melanogaster*, *C. elegans*, and *Arabidopsis thaliana* (DURET and MOUCHIROUD 1999), consistent with a model of weak selection on translational efficiency. Second, patterns of polymorphism for unpreferred synonymous mutations show a skew towards rarity in natural populations of *Drosophila simulans* in comparison with preferred substitutions (AKASHI 1997), indicating a role for weak purifying selection preventing their

spread. Third, preferred codons have been shown to be associated in some taxa with the most abundant tRNAs, as predicted by the translational efficiency model (BULMER 1987). Fourth, a correlation between synonymous and nonsynonymous substitution rate has been interpreted as reflecting stronger selection on codon usage in more highly constrained genes (COMERON and KREITMAN 1998; DUNN *et al.* 2001). Most recently, direct experimental manipulation of codon usage at the ADH locus in *Drosophila melanogaster* has shown a direct effect of the frequency of preferred codons on protein levels (CARLINI and STEPHAN 2003).

Substantial evidence is accumulating that codon bias is affected by population size and rates of crossing over. Several comparisons in Drosophila, including *D. simulans* vs. *D. melanogaster* (AKASHI 1996), and in Hawaiian Drosophila (KLIMAN *et al.* 2000), provide evidence for an accumulation of unpreferred codons in species with reduced effective size. Kliman and Hey (1993) first provided evidence for a relationship between rates of crossing over and average level of codon bias in Drosophila, and interpreted this as resulting from effects of hitchhiking. However, recent analyses which added the effect of base composition at noncoding sites suggested that at least some of the effect of recombination rate may reflect mutational biases in different regions (MARAIS *et al.* 2001), since most preferred codons in Drosophila are G- or C- ending, and there is a general increase in GC content in regions of high recombination. Alternatively, it has been hypothesized that this pattern is explained by the effects of biased gene conversion favouring the conversion of AT GC alleles (GALTIER *et al.* 2001). Assuming a good correlation between rates of this process and rates of crossing over, this could lead to an

elevated GC content in regions of high recombination (BIRDSELL 2002; MARAIS 2003).

Another approach to testing the influence of recombination on selection for codon usage bias involves examination of orthologous genes in related species where local rates of crossing over have changed. Takano-Shimizu (1999) compared levels of codon bias of several orthologous genes between *D. melanogaster* and *D. yakuba*, showing reduced codon bias in melanogaster, where recombination rates in the region were shown to be substantially reduced. However, further comparisons of codon bias across more distantly related Drosophila species also indicate that shifts in mutational biases may cause strong effects on the evolution of codon usage, perhaps stronger than effects of changes in effective size (RODRIGUEZ-TRELLES *et al.* 1999; TAKANO-SHIMIZU 2001).

Selection on noncoding DNA

A final, potentially highly significant source of weak selection is on the composition and size of non-coding regions of the genome, including introns and intergenic regions. These regions remain the most poorly understood components of genomes from a functional basis, but are clearly composed of regulatory regions, sequences that have structural importance, and repetitive DNA. In *Drosophila melanogaster* (CHARLESWORTH and LANGLEY 1989) and *Arabidopsis lyrata* (WRIGHT *et al.* 2001), transposable element (TE) insertions in noncoding regions appear to be segregating at lower frequency than expected under neutrality, consistent with a model of transposition-selection balance. Analysis of polymorphism data in humans has also provided support for the possible action of

selection on base composition in noncoding DNA of humans (EYRE-WALKER 1999; SMITH and EYRE-WALKER 2001). However, the action of selection on base composition is difficult to distinguish from biased gene conversion (GALTIER *et al.* 2001), since the processes will have the same effects on patterns of polymorphism and divergence, with the rate of biased gene conversion replacing the selection coefficient in population genetic models (LERCHER *et al.* 2002). TEs have been shown to accumulate in regions of low recombination in Drosophila, although this may reflect relaxation of natural selection in these regions if a major source of selection is the deleterious effects of ectopic recombination between elements (CHARLESWORTH and LANGLEY 1989). Additionally, observed correlations between rates of crossing over and both base composition in humans (FULLERTON *et al.* 2001), and intron size in Drosophila (CARVALHO and CLARK 1999; COMERON and KREITMAN 2000) have been interpreted as reflecting the action of weak purifying selection favouring small intron size and GC-biased composition in noncoding DNA. In all of these cases, rejecting neutral explanations for genomic distributions is difficult.

1.3 Effects of mating system on DNA polymorphism and divergence in plant populations

If weak purifying selection is frequent genome-wide, the mating systems of populations may have general effects on patterns of genome evolution. The mating system has major consequences for levels of genetic exchange and life history, which can strongly influence the structuring of genetic variability, the rate of evolution of weak deleterious and beneficial mutations, and thus the level of

molecular adaptation across genomes. Plants exhibit a wide variety of breeding systems, with close relatives exhibiting extreme contrasts in rates of self-fertilisation. As mutational biases and selection pressures are likely to be similar at orthologous loci in these close relatives, related inbred and outbred plants represent a useful system for studying the role of effective population size and recombination on the evolution of genes and genomes.

Under a strictly neutral model, increased levels of inbreeding reduce the effective size of a population as a function of the inbreeding coefficient (F) up to a maximum of a twofold reduction with complete self-fertilisation (NORDBORG 2000; POLLAK 1987). Furthermore, the effective rate of recombination is reduced in partially inbred populations, due to the rarity of double heterozygotes. This will lead to strong increases in linkage disequilibrium as a function of the selfing rate (Nordborg, 2000a). When effective recombination rates are low, strong purifying or directional selection will further reduce the effective population size of linked sites (CHARLESWORTH *et al.*, 1993). In the presence of natural selection at linked sites, the effective population size and genetic diversity of inbred populations can thus be further reduced (CHARLESWORTH *et al.* 1993), leading to a higher rate of fixation of slightly deleterious mutations (CHARLESWORTH 1994), and a lower chance of fixation of beneficial mutations. With very high levels of self-fertilisation, inbreeders may also be susceptible to the effects of Muller's ratchet, the stochastic loss of individuals free from deleterious mutations (CHARLESWORTH 1993; HELLER and MAYNARD SMITH 1979). Finally, inbreeding plant populations may often have a 'weedy' life history associated with strong population structure, a high level of variability in population size, low levels of migration and frequent extinction and re-

colonization. These demographic effects can cause further reductions in effective population size (WHITLOCK and BARTON 1997) and neutral variability (PANNELL and CHARLESWORTH 2000). Substantial data on levels of allozyme polymorphism (Hamrick and Godt, 1990), as well as recent work on nucleotide variability in Leavenworthia (LIU *et al.* 1999; LIU *et al.* 1998), Lycopersicon (BAUDRY *et al.* 2001) and Arabidopsis (BERGELSON *et al.* 1998; SAVOLAINEN *et al.* 2000), have provided evidence for a strong reduction in diversity within selfing populations, and greater differentiation between populations. The relative importance of effects of demography and selection remain unclear, and are likely to require further information on patterns of variability within and between populations. Regardless of the dominant causes, in the face of highly reduced effective population sizes, the potential for the accumulation of weak deleterious mutations and the difficulty of fixing beneficial mutations has led to the hypothesis that self-fertilization is an evolutionary 'dead-end' that leads to eventual extinction, although the phylogenetic support for this hypothesis remains weak (TAKEBAYASHI and MORRELL 2001).

While there is a general prediction that the efficacy of natural selection should be reduced as a function of the rate of self-fertilization, several complications arise when attempting to assess the quantitative effects of the mating system on genome evolution. First, with partially recessive effects of mutations, the strength of selection may be effectively increased in selfing populations, due to higher levels of homozygosity. However, the effect of reduced N_e is likely to be stronger than the increased selection, since the rate of fixation of deleterious mutations is in general only weakly affected by the dominance coefficient (CHARLESWORTH 1994). A second complication arises with the presence of population structure. The models of

hitchhiking discussed above predict strong reductions in within-population effective population size. However, with strongly subdivided populations, effective sizes might remain high species-wide (CHARLESWORTH *et al.* 1997). In fact, in situations where multiple populations have been studied, selfers appear to maintain substantial polymorphism species-wide, perhaps maintaining values as high as those found in related outcrossers (LIU *et al.* 1999; LIU *et al.* 1998; SAVOLAINEN *et al.* 2000). Thirdly, the effect of selfing rate on the relationship between crossing over and the efficacy of selection is also unclear. Because recombination is suppressed genome-wide in inbreeders, the relationship between crossing over and molecular adaptation may break down, since they will effectively cover a much lower range of recombination rates than related outcrossers. If so, heterogeneity across selfing genomes may primarily reflect mutational biases. Indeed, recent evidence on the distribution of transposable elements (DURET *et al.* 2000) and codon bias (MARAIS *et al.* 2001) in the self-fertilising *C. elegans* appears consistent with purely mutational explanations. While this was interpreted as evidence against hitchhiking models, it may in fact be an inevitable consequence of hitchhiking in inbred organisms. Finally, if biased gene conversion contributes substantially to patterns of base composition evolution, the mating system of populations may be expected to be generally associated with neutral genomic changes. Since biased gene conversion will only occur at heterozygous sites, the rate of this process will be dramatically reduced in highly inbred populations. This could lead to strong differences in nucleotide composition among species, regardless of differences in the strength of selection. Clearly, the interaction of mating system, recombination rate and population

structure on patterns of polymorphism and divergence at both neutral and selected sites requires both more empirical investigation and theoretical understanding.

1.4 Molecular population genetics and genomics in Arabidopsis

Arabidopsis thaliana is a highly self-fertilising (ABBOTT 1989) weedy annual, and is the first plant for which the complete genome sequence was sequenced. The *A. thaliana* genome is small, with a high gene density and relatively low levels of repetitive DNA (ARABIDOPSIS GENOME INITIATIVE 2000). The species has a worldwide distribution, although it is thought to have colonized North America recently. Recent phylogenetic work indicates that the closest relatives of *A. thaliana* include the self-incompatible, highly outcrossing species *Arabidopsis lyrata* and *Arabidopsis halleri* (KOCH *et al.* 2001; KOCH *et al.* 2000). Since the divergence from these species, the chromosome number appears to have been reduced from 8 to 5 chromosomes in the *A. thaliana* lineage. *A. lyrata* is a circumpolar species, with a subspecies found in Europe (ssp. *petrea*) and ssp. *lyrata* distributed in the Northern U.S., Canada, and Japan.

DNA sequence polymorphism has now been studied for a number of *A. thaliana* loci, usually using samples of single individuals from different populations, and several general observations have been made. First, although average within-population diversity is low for all types of markers studied (ABBOTT 1989; BERGE *et al.* 1998; BERGELSON *et al.* 1998; TODOKORO *et al.* 1995), species-wide polymorphism is high; average silent diversity is only slightly lower than in *Drosophila melanogaster*. Recent analysis of several large North American populations also suggests that some populations may contain a significant fraction of

the species-wide diversity (E.A. Stahl, R.A. Hufft, M. Kreitman, and J. Bergelson, submitted manuscript).

Second, the frequency distribution of polymorphisms at many loci in species-wide samples is skewed towards rare variants in *A. thaliana*; this is detectable by Tajima's D statistic (e.g. PURUGGANAN and SUDDITH 1999; KUITTINEN and AGUADE 2000; KAWABE *et al.* 2000). This observation, combined with the slight evidence for genetic isolation by geographic distance (BERGELSON *et al.* 1998), led to suggestions of rapid population expansion (INNAN and STEPHAN 2000; KUITTINEN and AGUADE 2000). Recently, however, AFLP studies have provided evidence for a weak but significant relationship between geographic and genetic distance in Eurasian populations (SHARBEL *et al.* 2000). At least some of the signal of population expansion in species-wide samples may thus result from recent colonization of North America by *A. thaliana*. Thirdly, many *A. thaliana* loci show an excess of amino acid replacement polymorphisms compared with fixed differences from other species, suggesting that unusually many slightly deleterious mutations segregate in natural populations (KAWABE *et al.* 1997; PURUGGANAN and SUDDITH 1998; PURUGGANAN and SUDDITH 1999). Fourth, nucleotide diversity studies at several loci suggest the presence of two major classes of haplotypes, with linkage disequilibrium between different polymorphic sites, although there is also clear evidence for recombination within loci (INNAN *et al.* 1996; KUITTINEN and AGUADE 2000). The haplotype structure has been interpreted as possible evidence for long-term population structure or balancing selection (KAWABE and MIYASHITA 1999). However, the neutral coalescent process with low levels of recombination will frequently generate two major haplotypes (HUDSON 1990), and it is not clear that *A.*

thaliana shows any general excess of haplotype structure compared with the appropriate neutral model taking selfing into account (AGUADE 2001). Finally, recent studies of polymorphism across several large genomic regions have shown linkage disequilibrium extending over more than one hundred kilobases in *A. thaliana* (HAGENBLAD and NORDBORG 2002; NORDBORG *et al.* 2002). This is much more extended linkage disequilibrium than is found in *D. melanogaster*, though the two species have comparable polymorphism levels.

To date, levels of within-population nucleotide diversity have been compared between *A. thaliana* and *A. lyrata* at the ADH1 locus (BERGELSON *et al.* 1998; SAVOLAINEN *et al.* 2000). While this study provided preliminary evidence for higher levels of within-population diversity in *A. lyrata* and lower levels of species-wide amino acid polymorphism, evidence for significant departures from neutrality at this locus in the primary population studied in *A. lyrata* indicates the importance of examining patterns of nucleotide diversity at multiple loci using larger population samples, to obtain a better assessment of levels and structuring of genetic variability across the genome and across populations.

In addition to preliminary suggestions of reduced efficacy of selection on amino acid mutations in *A. thaliana*, there may also be a reduction in selection against silent sites. The correlation between gene expression and codon bias in *A. thaliana* has provided evidence for selection on synonymous sites, but the slope of this relationship, and the average codon bias is much smaller than is observed in Drosophila (DURET and MOUCHIROUD 1999). While this broad-scale difference is consistent with effects of breeding system on the efficacy of selection, comparisons of codon bias with related outcrossing species are required to test this.

1.5 Thesis outline

In this thesis, I investigate the consequences of recombination and inbreeding on patterns of nucleotide polymorphism and molecular evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*. I analyse nucleotide polymorphism and molecular evolution in both species, and genome-wide structure in *A. thaliana* to test for the effects of mating system and recombination rate on genome diversity and evolution. In **Chapter 2** I compare codon bias, intron size and rates of protein evolution between *A. thaliana* and *A. lyrata* to test for a reduction in the efficacy of selection in the selfing *A. thaliana*. The molecular evolution data are also used to estimate the genomic deleterious mutation rate of amino acids in these species, to investigate the potential for high levels of background selection in *A. thaliana*. In **Chapter 3** I incorporate both polymorphism and divergence data to compare the action of natural selection on codon bias and amino acid variants between the two species. I also make use of recent information on tRNA gene abundance to test models of translational selection for the evolution of codon bias. In **Chapter 4** I study nucleotide polymorphism levels at five nuclear genes and one chloroplast gene, in four populations of *Arabidopsis lyrata* and a species-wide sample from *Arabdiopsis thaliana*. I compare both levels of diversity and patterns of linkage disequilibrium, to test equilibrium models predicting reductions in both diversity and effective recombination rates in inbreeders. **Chapter 5** presents analysis of the complete genome of *A. thaliana* to investigate the relative importance of recombination rate and gene density in driving transposable element distributions in this species. I use this analysis to investigate the prediction that differences in recombination rates

across the genome of a self-fertilizing species are less important in driving differences in the efficacy of natural selection than differences in gene density. In **Chapter 6** I investigate the relationship between recombination rate and nucleotide diversity in *Arabidopsis lyrata* using 18 loci that map to the first chromosome of *A. thaliana*, to test for the action of genetic hitchhiking in regions of reduced recombination in this outcrossing species. In **Chapter 7** I summarize the major findings, use the combined results to make some new inferences, and discuss future directions.

Chapter 2

Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*

2.1 Introduction

The evolution of self-fertilization is associated with a large reduction in the effective rate of recombination, and a corresponding decline in effective population size. If many spontaneous mutations are slightly deleterious, this shift in breeding system is expected to lead to a reduced efficacy of natural selection, and genome-wide changes in the rates of molecular evolution. Here, we investigate the effects of breeding system on molecular evolution in the highly self-fertilizing plant *Arabidopsis thaliana*, by comparing coding and noncoding genomic regions with those of its close outcrossing relative, the self-incompatible *Arabidopsis lyrata*.

We use two main approaches to compare the effectiveness of selection: a) pairwise comparison of codon usage bias and intron length for genes sequenced in both species, and b) comparison of the relative rates of nucleotide substitution for different classes of mutation, using more distantly related species in the Brassicaceae to polarize the substitutions along each lineage. We also utilize genomic information on gene expression to evaluate the influence of expression level on selection coefficients across the sampled loci. The extent to which the efficacy of natural selection is reduced in inbreeding species may depend importantly on the genomic rate of deleterious mutations (U) (CHARLESWORTH 1994). We therefore also use our analysis to estimate U for amino acid substitutions,

using the relative rates of synonymous and nonsynonymous mutations between species (KEIGHTLEY and EYRE-WALKER 2000).

2.2 Methods

Sequence information

Table 2.1 shows the genes studied, the species from which they were available, and the source of the sequences. Most sequences were partial coding regions, including multiple exons and introns. The sample includes twenty-three nuclear loci distributed across all five *A. thaliana* chromosomes, and a single locus (*matK*) from the chloroplast genome. Sequences from *A. thaliana* were extracted from Genbank (National Center for Biotechnology Information, NCBI; http://www.ncbi.nlm.nih.gov), using either the complete genome sequence from the Columbia ecotype, or else a sequence from a published population survey. All nuclear genes previously sequenced in *A. lyrata* were extracted from Genbank, along with their orthologs in *A. thaliana*, with the exception of the putative self-incompatibility locus *SRK*, which is likely to be a pseudogene in A. *thaliana* (KUSABA *et al.* 2001; SCHIERUP *et al.* 2001). The *A. lyrata* sequences are derived from different source populations, including representatives from the European subspecies *petraea*, the Japanese subspecies *kawasakiana* and the North American subspecies *lyrata*. Seven additional single-copy loci were selected for sequencing in *A. lyrata* (see below). All *A. lyrata* loci were submitted to a BLAST search (ALTSCHUL *et al.* 1990; ALTSCHUL *et al.* 1997); http://www.ncbi.nlm.nih.gov/BLAST), to confirm that the locus has a single clear ortholog in *A. thaliana*, and to identify sequences available

Table 2.1 Genes surveyed in analysis of molecular evolution in *A. thaliana* and *A. lyrata*. Locus names are based on *A. thaliana* genome project.

Locus	Description	Species	Source[1], Accession number	Relative EST abundance
ABAp1	putative ABA binding protein	*A. thaliana*	AT4G01600	0
		A. lyrata	this study, AF494370	
ABC1At	ABC transporter	*A. thaliana*	AT4G01660	51.41
		A. lyrata	this study, AF494371 (5'), AF494372(3')	
		C. rubella	this study, AF494373	
ADH1	alcohol dehydrogenase 1	*A. thaliana*	AT1G77120	33.33
		A. lyrata	AF110453	
		C. rubella	AF110435	
AOP1	2-oxoglutarate-dependent dioxygenase 1	*A. thaliana*	AT4G03070	0
		A. lyrata	AF417857	
AOP2	2-oxoglutarate-dependent dioxygenase 2	*A. thaliana*	AF417858	0
		A. lyrata	(AT4G03060)	
		Brassica oleraceae	AF418239 AY044425	
AOP3	2-oxoglutarate-dependent dioxygenase 3	*A. thaliana*	AF417859	2.55
		A. lyrata	(AT4G03050) AF418280	
AP1	apetala 1	*A. thaliana*	AT1G69120	7.65
		A. lyrata	AF143379	
		B.oleraceae	Z37968	
AP3	apetala 3	*A. thaliana*	AT3G54340	14.07
		A. lyrata	AF143380	
		B. oleraceae	U67456	
CAUL	Cauliflower	*A. thaliana*	AT1G26310	2.55
		A. lyrata	AF143381	
		B. rapa	AJ251300	
CHI	chalcone flavone isomerase	*A. thaliana*	AT3G55120	5.10
		A. lyrata	AJ287322	
		R. sativus	AF031921	
ChiA	acidic endochitinase	*A. thaliana*	AT5G24090	7.65
		A. lyrata	AB006072	
		A. glabra	AB006071	
CHS	chalcone synthase	*A. thaliana*	AT5G13930	98.49
		A. lyrata	AF112100	
		C. rubella	AF112106	
EnCoH1	enoyl-CoA hydratase	*A. thaliana*	AT4G16800	0
		A. lyrata	this study, AF494369	
		C. rubella	AJ400821	
ETR1	ethylene receptor 1	*A. thaliana*	AT1G66340	5.10
		A. lyrata	this study, AF494374	
		B. oleraceae	AF047476	
F3H	flavanone-3-hydroxylase	*A. thaliana*	AT3G51240	16.67
		A. lyrata	AJ295607	
		M. incana	X72594	
FAH1	ferulate-5-hydroxylase	*A. thaliana*	AT4G36220	5.10
		A. lyrata	AJ295586	
		B. napus	AF214007	

Table 2.1, continued

FKA1	fructokinase 1	*A. thaliana*	AT2G31390	32.27
		A. lyrata	this study, AF494374	
GL1	glabrous 1	*A. thaliana*	AT3G27920	0
		A. lyrata	AF263720	
HAT4	homeobox protein 4	*A. thaliana*	AT4G16780	12.76
		A. lyrata	this study, AF494367	
		C. rubella	AJ400821	
matK	maturase K	*A. thaliana*	Chlor.- Pos. 2056	-
		A. lyrata	AF144342	
		A. glabra	AF144333	
PGIC	cytosolic phosphoglucose isomerase	*A. thaliana*	AT5G42740	8.33
		A. lyrata	AB044969	
		L. crassa	AF054455	
PIST	Pistillata	*A. thaliana*	AT5G20240	42.21
		A. lyrata	AF143382	
RPM1	disease resistance protein	*A. thaliana*	AT3G07040	5.10
		A. lyrata	AF122982	
		B. napus	AF105139	
Sc-ADH	Short-chain alcohol dehydrogenase	*A . thaliana*	AT4G05530	15.31
		A. lyrata	this study, AF494368	

[1], for *A. thaliana*, locus names from the complete genome sequence are given. In cases where the sequence was derived from a population study, references and accession numbers are given, with locus names in parentheses.

from the closest outgroup species in the Brassicaceae (KOCH *et al.* 2001; KOCH *et al.* 2000). Two of the loci sequenced for this study (*HAT4* and *EnCoH1*) were selected from a 28kb genomic region of the *A. thaliana* chromosome 4 that has been sequenced in the outgroup species *Capsella rubella* (ACARKAN *et al.* 2000), which belongs to the sister group to *A. thaliana* and *A. lyrata* (KOCH *et al.* 2001; KOCH *et al.* 2000). For an additional locus, *ABC1At*, the orthologous region was sequenced in *C. rubella*, using DNA provided by H. Hurka. In total, seven loci in the analysis have an outgroup sequence from this sister group, including *C. rubella* and *A. glabra*. For an additional eight genes, sequences from more diverged outgroup species from the genera Brassica, Leavenworthia, Matthiola or Raphanus were used. Because single sequences from each species are used in the comparison, estimates of substitution rates along each lineage will include segregating polymorphisms as well

as fixed differences between species. However, this should not bias the results, provided that the species-wide coalescence times are similar in the two species. Given that species-wide estimates of silent polymorphism do not appear to be very different in *A. thaliana* and *A. lyrata* (SAVOLAINEN *et al.* 2000) (despite highly reduced within-population variability in *A. thaliana*), this is reasonable in the present case.

Table 2.1 also shows for each locus the maximum number of matches to expressed sequence tags (ESTs) from the *Arabidopsis thaliana* EST projects available in Genbank, obtained from G. Marais (unpublished data). This index has been normalized for differences in total EST number across the libraries, by dividing each value by the total number of ESTs in the source library. The number of EST matches is often used as a measure of expression level, and correlates with levels of codon bias in *A. thaliana* (DURET and MOUCHIROUD 1999).

DNA extraction, gene amplification and sequencing

DNA was extracted from *A. lyrata* individuals using the CTAB protocol (JUNGHANS and METZLAFF 1990). PCR primers were designed using the *A. thaliana* sequence information, and amplified in *A. lyrata* for 30 cycles consisting of 1 min. denaturing at 95^0C, 30 seconds annealing at 55^0C, and 2 min. extension at 72^0C. PCR primer sequences are: Sc-ADH-F 5'GGCATTCCTCCAGCGAC3', Sc-ADH-R 5'CTTCCGTCGTCGTCTCTTC3'; EnCoA-F 5'CTGGTCGGTTACTTTTGTCG3', EnCoA-R 5'CCTGTCACCAAAAATGCTATT3'; HAT4-F 5'CGTGAACAGACCACCGTC3', HAT4-R 5'AGCGTCAAAAGTCAAGCCGT3'; ABAp1-F

5'CAAGCACAAAACCAACAGCC3', ABAp1-R 5'CAAACCCATCTCGTGTCACC3'; ABC1At-F 5'CTTTACCAGGCTCGTCAATG3', ABC1At-R 5'CATCACATCAGCACCTTGAC3'; ETR1-F 5'-AGACCAAAGCTCATGCATTTCT-3', ETR1-R 5'-TGTTGACTCATGAGATTAGAAGCA-3'; FKA1-F 5'-CCTGGATTCCTCAAAGCTCC-3', FKA1-R 5'-TCCCAAATGCTCATGATCTG-3'. PCR fragments were cloned into the PCR 2.1 vector using the TA cloning kit (Invitrogen Life Technologies), and at least five clones were sequenced using the ABI automated sequencing facilities at the University of Edinburgh Institute of Cell, Animal and Population Biology. PCR primers, the Universal M13 primers, as well as several internal sequencing primers were used in sequencing. In all cases, sequence analysis of clones indicated that the genes were single copy in *A. lyrata*, with one or two haplotypes identified per individual, with the exception of several clear PCR recombinants.

Rates of protein evolution

Single sequences from *A. thaliana*, *A. lyrata* and, where available, an outgroup species, were aligned using the CLUSTALW computer package (THOMPSON *et al.* 1994), and alignments were subsequently corrected by eye using the sequence editor GENEDOC (NICHOLAS 1997). Pairwise estimates of K_a, the number of nonsynonymous substitutions per site, and K_s, the number of synonymous substitutions per site, were calculated for each gene using the program K-estimator v. 5.5 (COMERON 1999), which uses the method of Comeron (1995) to estimate

substitution rates. K-estimator was also used to obtain 95% confidence intervals for these estimates by Monte Carlo simulation. Total pairwise substitution rates were estimated by combining the sequences of all nuclear loci.

When an outgroup sequence was available, two approaches were utilised to estimate rates of synonymous and nonsynonymous substitutions in *A. thaliana* and *A. lyrata* independently. First, parsimony was used to estimate directly the numbers of nonsynonymous and synonymous substitutions in both lineages (AKASHI 1996). For some sites, multiple substitutions precluded inference based on parsimony, and these were excluded. For parsimony analysis, synonymous substitutions were counted only for sites that did not have a nonsynonymous difference. Differences between the lineages in total substitution rates were assessed using Tajima's relative rate test (TAJIMA 1993). Second, rates of synonymous and nonsynonymous substitutions per site were estimated using the maximum likelihood method of the CODEML program, in the PAML computer package (YANG 1997). This program estimates substitution rates, taking into account multiple substitutions per site, different rates of transitions and transversions, and effects of codon usage. Two models of sequence evolution were considered: 1) a model with a fixed K_a/K_s ratio across lineages, and 2) a model that allowed this ratio to differ for each species. Significance was assessed using the chi-square test with two degrees of freedom, where the chi square statistic is $2(L_2-L_1)$, L_2 is the log likelihood for the second (free ratios) model, while L_1 is the log likelihood for the model with fixed ratios (see Yang, 1998). Standard errors were also estimated for the K_a/K_s ratios along each lineage using the PAML program, although these estimates provide only an approximate description of the likelihood surface (YANG 1997).

Comparisons of codon bias

Levels of codon bias and patterns of codon usage were examined for each locus in both *Arabidopsis* species. The GC content at third codon positions (GC_3) was used to measure codon usage bias. Chiapello et al. (1998) have shown that in *A. thaliana* GC_3 is highly correlated with the degree of biased codon usage, and with gene expression levels. This measure is preferable to another standard measure of codon bias, ENC, since it measures more directly the frequency of preferred codons. Pairwise comparisons were also made between *A. lyrata* and *A. thaliana* for the presence of major vs. minor codons, as defined by multivariate analysis of codon usage in *A. thaliana* (CHIAPELLO *et al.* 1998). In particular, for all codons which have the same amino acid, the number of cases where *A. lyrata* has a major codon and *A. thaliana* a nonmajor codon, and vice versa, were recorded (AKASHI 1996). With the outgroup sequences, rates of unpreferred relative to preferred synonymous substitutions were also estimated for each lineage independently, using the assumptions of parsimony (AKASHI 1995; AKASHI 1996; TAKANO-SHIMIZU 1999). Only codons that encode the same amino acid in all three lineages were used in this analysis. Because of the small number of *A. lyrata* genes currently available, this analysis assumes that codon preferences are the same as those in *A. thaliana*.

Estimating the genomic deleterious mutation rate (U)

The combined sequence dataset of all nuclear loci in *A. thaliana* and *A. lyrata* was also used to estimate *U*, the genomic deleterious mutation rate for *Arabidopsis*, using the method of Keightley and Eyre-Walker (2000), which considers only the contribution of amino acid changes to deleterious mutation. Because this method uses estimates of substitution rates between a pair of species, it measures the average deleterious mutation rate for both species. The per-site estimate of the number of deleterious mutations between species (u) was calculated from the following equation using a program provided by P. Keightley (pers. com.):

$$u = (\overline{K_{ts}N_{ts}} + \overline{K_{tv}N_{tv}}) - \frac{\overline{K_n}}{3}$$

where K_{ts} is the per-codon number of synonymous transitions, K_{tv} is the per-codon number of synonymous transversions, K_n is the per-codon nonsynonymous substitution rate and N_{ts} and N_{tv} are estimates of the proportion of the transitions and transversions in the sequence that would cause an amino acid substitution. To convert this into an estimate of the genome wide deleterious mutation rate, the per-site value needs to be multiplied by the quantity Z:

$$Z = \frac{2 \times S \times I}{T}$$

where S is the total number of base pairs of exon sequence in the genome, I is the generation time, and T is the number of years of divergence, or twice the divergence time between the species. S was estimated using information from the *A. thaliana* genome sequence project (Arabidopsis Genome Consortium, 2000: S=33,249,250). I

is thought to be 1 generation/year or less in natural populations of *A. thaliana*, while it varies from 1-2 for *A. lyrata*. For the calculation of U, I was assumed to be 1. The divergence time between *A. thaliana* and *A. lyrata* is unknown, and few estimates of silent substitution rates per unit time exist for dicotyledonous plants. Koch et al. (2000) give an estimate of silent substitution rates as 1.5×10^{-8} per year for the Brassicaceae, citing a study of fossil pollen deposits. This estimate of nuclear substitution rate is at the high end of those made across diverse plants, so we also use a lower-bound estimate of the per year substitution rate of 5.8×10^{-9} by Wolfe *et al.* (1987), which uses the divergence between monocots and dicots. Using these two substitution rates, we estimated T from the total number of synonymous substitutions per synonymous site between the two species from our complete dataset of nuclear loci.

2.3 Results

2.3.1 Rates of protein evolution

Table 2.2 summarizes pairwise estimates of synonymous and nonsynonymous divergence between *A. thaliana* and *A. lyrata* for the 24 loci. Levels of selective constraint, as measured by K_a/K_s ratios, are highly variable across loci, although the total ratio for nuclear loci is 0.17, indicating significant purifying selection on amino acid substitutions. However, five loci show K_a/K_s ratios greater than 0.4, suggesting that their protein sequences are evolving rapidly. One locus, *ABC1At*, has accumulated numerous deletions and frameshifts in its 5' end in *A. lyrata*, and is thus likely to be a pseudogene in this species. Within the region surveyed, there are at least five large deletions in *A. lyrata*, including three within

exons. This is surprising, given evidence for the essential function and high expression of this protein in *A. thaliana* (CARDAZZO *et al.* 1998). Sequencing of this region in *C. rubella* confirmed that these deletions are specific to *A. lyrata*, and after elimination of the deleted regions, parsimony analysis suggests a large excess of replacement substitutions in *A. lyrata* (Table 2.2), although frameshifts preclude an accurate estimate. To examine the evolution of this gene in more detail, we subsequently amplified and sequenced the 3' portion of this locus. In sharp contrast to the 5' end, this region had no deletions, and relative rates of replacement and synonymous substitution suggest strong selective constraints (Table 2.2). This indicates that the gene has either become truncated, or, more likely, that it has been duplicated, with the second copy having degenerated. Because the sequenced 5' end of this locus appears to be evolving neutrally in *A. lyrata*, this region of the gene was excluded from global comparisons of relative rates of substitution between the species.

Two additional nuclear loci, *AOP3* and *ABAP1*, show very high K_a/K_s ratios. Although this appears to be in part due to low estimates of K_s for these loci, *AOP3* has a relatively high K_a, along with the other members of the *AOP* gene family. Relaxed constraint on these genes is perhaps not unexpected, given evidence that these loci involved in glucosinolate production are expressed only in some populations of *A. thaliana* (KLIEBENSTEIN *et al.* 2001).

The single locus sequenced from the chloroplast genome, *matK*, also shows an unusually high K_a/K_s ratio, which is consistent with previous studies suggesting that it has one of the lowest levels of constraint among chloroplast genes, despite being widely distributed among plants (YOUNG and DEPAMPHILIS 2000). *MatK* also

shows low synonymous divergence compared with the nuclear loci, which is in accordance with the previously estimated greater than twofold reduction in synonymous substitution rates in chloroplast genes (WOLFE *et al.* 1987).

Despite a generally high level of selective constraint (STAHL *et al.* 1999), the coding region of the disease resistance locus *RPM1* contains a region with a deletion in *A. lyrata* compared with both *A. thaliana* and the outgroup sequence, leading to a large number of amino acid replacements in this region, which was therefore excluded from estimates of substitution rate. The *GLB1* locus, which is well characterized in *A. thaliana*, differs in *A. lyrata* by a frameshift in the last exon, leading to a smaller protein (HAUSER *et al.* 2001), and this region was also excluded from subsequent analysis.

With the exception of the putative *ABC1At* pseudogene, the genes with highest K_a/K_s ratios tend to be those that have few or no matches to ESTs, suggesting that they are expressed at low levels (Table 2.1). Indeed, a strong negative correlation is observed between the maximum number of EST matches and K_a/K_s (Spearman r=-0.671, p<0.01). The correlation appears to primarily reflect fewer nonsynonymous substitutions in highly expressed genes, rather than excess synonymous substitutions (K_a: Spearman r=-0.568, p<0.05; K_s: Spearman r=-0.303, p>0.05). This effect might reflect greater selective constraint on genes that are more broadly or highly expressed. Alternatively, it may be due to a higher level of annotation error for low-expression genes, which are generally less well characterized. For example, if the prediction of exon positions is generally poorer for genes that are less expressed, estimates of the K_a/K_s ratio could be inflated. However, this does not appear to be the cause of the effect; using only genes that

have a complete cDNA sequenced in *A. thaliana*, the effect of EST number on K_a/K_s remains highly significant (Spearman r=-0.668, p<0.01).

Table 2.2 also shows parsimony-based estimates of synonymous and nonsynonymous substitutions in *A. thaliana* and *A. lyrata* for each locus with an available outgroup sequence. These analyses show that substitutions occur equally along both branches; there is no significant difference between the species in either synonymous (Tajima test χ^2=1.40, p>0.05) or nonsynonymous (Tajima test χ^2=0.36, p>0.05) substitution rates. Similarly, there is no consistent difference in level of constraint between the species; the ratio of replacement to synonymous substitutions is not significantly different between species (G=0.021, p>0.05). Analysing genes with low (K_a/K_s>0.1, N=7) and high (K_a/K_s>0.1, N=9) constraint separately, no significant difference in the ratio of nonsynonymous to synonymous substitutions is observed between the species for either class of genes (low, G=0.292, p>0.05; high, G=0.088, p>0.05).

Some of the outgroup species have high levels of silent divergence from *A. thaliana* and *A. lyrata*, making it difficult to infer the lineage in which substitutions have occurred (BROMHAM *et al.* 2000; TAJIMA 1993). As the relative distance of ingroup to outgroup increases, the power of the relative rate test is further decreased (BROMHAM *et al.* 2000). If the analysis is restricted to the six loci with outgroup sequences from *C. rubella* or *A. glabra*, which belong to the sister group to *A. thaliana* and *A. lyrata*, we still observe no lineage effects for either synonymous or nonsynonymous rates, or for K_a/K_s (p>0.05). However, the number of synonymous

Table 2.2 Synonymous and nonsynonymous substitution rates between *A. thaliana* and *A. lyrata*. Rates were estimated using the method of (COMERON 1999).

locus	S^1	R^2	K_s (95% C.I.)[3]	K_a (95% C.I.)[3]	K_a/K_s	Syn_{thal}[4]	Rep_{thal}[4]	Syn_{lyr}[4]	Rep_{lyr}[4]
ABAp1	102.3	323.3	0.051 (0.007-0.107)	0.029 (0.010-0.051)	0.569	-	-	-	-
ABC1At (5')	67.1	178.4	0.168 (0.062-0.304)	0.0967 (0.0482-0.152)	0.575	3	0	7	14
ABC1At (3')	188.8	497.8	0.107 (0.058-0.166)	0.0020 (0-0.006)	0.019	-	-	-	-
ADH	317.6	859.4	0.147 (0.099-0.199)	0.0195 (0.0098-0.0288)	0.133	19	10	18	4
AOP1	223.0	737.3	0.229 (0.157-0.317)	0.0916 (0.0648-0.1281)	0.400	-	-	-	-
AOP2	303	892	0.0944 (0.054-0.139)	0.0379 (0.0417-0.0763)	0.402	10	10	2	10
AOP3	289.5	869.5	0.0953 (0.045-0.123)	0.0583 (0.0379-0.0728)	0.568	-	-	-	-
AP1	220.8	605.4	0.0730 (0.031-0.128)	0.0133 (0.0038-0.0245)	0.182	6	4	7	3
AP3	173.3	575.0	0.0854 (0.041-0.149)	0.0135 (0.0043-0.0271)	0.158	5	3	5	3
CAUL	196.3	583.5	0.0987 (0.048-0.156)	0.0315 (0.0170-0.0523)	0.319	7	7	8	8
CHI	179.8	534.8	0.210 (0.139-0.297)	0.0461 (0.0211-0.0772)	0.219	13	6	9	6
CHIA	226.0	678.0	0.114 (0.069-0.168)	0.0348 (0.0211-0.0502)	0.304	15	10	7	10
CHS	318.0	872.0	0.149 (0.098-0.205)	0.0075 (0.0022-0.0141)	0.050	19	2	19	2
EnCoH1	181.7	532.0	0.139 (0.080-0.211)	0.0076 (0.0018-0.0173)	0.054	14	2	9	1
ETR1	434.0	1150.1	0.150 (0.103-0.198)	0.0052 (0.0009-0.0096)	0.035	17	2	27	3
F3H	271.5	858.7	0.188 (0.129-0.257)	0.0070 (0.0013-0.0142)	0.037	17	3	19	3

Table 2.2, continued

locus	S[1]	R[2]	K_s (95% C.I.)[3]	K_a (95% C.I.)[3]	K_a/K_s	Syn_{thal}[4]	Rep_{thal}[4]	Syn_{lyr}[4]	Rep_{lyr}[4]
FAH1	362.1	1047.0	0.110 (0.069-0.156)	0.0086 (0.0030-0.0147)	0.078	15	4	15	3
FKA1	113.6	319.7	0.1631 (0.066-0.224)	0	0	-	-	-	-
GL1	168.2	484.3	0.0561 (0.015-0.106)	0.0146 (0.0046-0.0286)	0.260	-	-	-	-
HAT4	106.5	274.3	0.142 (0.062-0.254)	0.0037 (0-0.0112)	0.026	9	0	5	1
matK	362.8	1166.1	0.0267 (0.010-0.049)	0.0161 (0.0078-0.0237)	0.602	5	7	4	9
PGIC	397.0	1316.0	0.1570 (0.116-0.202)	0.0084 (0.0035-0.0146)	0.054	23	5	19	3
PIST	186.1	504.8	0.0978 (0.041-0.173)	0.0211 (0.0078-0.0347)	0.216	-	-	-	-
RPM1	761.1	2117.5	0.1011 (0.073-0.131)	0.0206 (0.0145-0.0276)	0.161	23	17	20	15
Sc-ADH	140.4	393.1	0.1501 (0.079-0.235)	0.0051 (0-0.015)	0.034	-	-	-	-
total	5776[5]	16950[5]	0.126[5]	0.0211[5]	0.167[5]	217[6]	92[6]	193[6]	84[6]

[1], number of synonymous sites
[2], number of replacement sites
[3], 95% confidence intervals
[4], parsimony-based estimates of number of synonymous (Syn) and replacement (Rep) changes, in *A. thaliana* (*thal*) and *A. lyrata* (*lyr*) (see Methods for details)
[5], total values excluding the *matK* chloroplast locus, and the putative *ABC1At* pseudogene.
[6], excludes only the putative *ABC1At* pseudogene.

substitutions in this sample is consistently greater in *A. thaliana*; 5 genes have higher numbers in *A. thaliana*, while no genes have more substitutions in *A. lyrata* (Wilcoxon signed ranks Z=-2.03, p<0.05).

Maximum likelihood estimates of substitution rates, which take substitution biases and multiple substitutions per site into account, are broadly consistent with the parsimony-based analysis. Table 2.3 shows that the estimated K_a/K_s ratios for each locus are similar in both the *A. thaliana* and *A. lyrata* lineages, with no consistent difference in level of selective constraint. For the vast majority of genes,

there was no evidence for a departure from a fixed level of selective constraint across all lineages, providing no evidence for a consistent change in selective constraint since the divergence of these two lineages from the outgroup species. Only one locus, *AOP2*, shows a significant lineage effect on K_a/K_s, although this would not be significant after correcting for multiple tests. In this case, the estimated K_a/K_s ratio in *A. lyrata* is strikingly higher than 1, but this appears to be largely due to an unusually low estimated synonymous substitution rate (Table 2.2). For this locus, a model which allows the *A. thaliana* ratio to vary, while the outgroup and *A. lyrata* have the same ratio, differs significantly from a fixed ratio model (χ^2=7.24, 1d.f., p<0.05), while a model allowing a free ratio in *A. lyrata* does not significantly improve the likelihood (χ^2=1.11, p>0.05). This suggests that the lineage difference at this locus largely represents reduced K_a/K_s in *A. thaliana*.

Table 2.3. Maximum likelihood estimates of the ratio of nonsynonymous to synonymous substitutions in *A. lyrata* and *A. thaliana*.

locus	K_a/K_s (S.E.)		2(L1-L2)[1]
	A. thaliana	*A. lyrata*	
ADH	0.1592 (0.0635)	0.0802 (0.0419)	4.28
AOP2	0.15 (0)	89.0 (31.46)	10.46*
AP1	0.1368 (0)	0.1544 (0.0679)	0.67
AP3	0.1462 (0.0963)	0.1486 (0.104)	0.012
CAUL	0.2962 (0.107)	0.2393 (0.435)	0.64
CHI	0.2574 (0.00524)	0.1973 (0.0836)	2.89
CHIA	0.2154 (0.0945)	0.3915 (0.128)	4.84
CHS	0.0278 (0.0211)	0.0545 (0.0316)	2.64
EnCoH1	0.0376 (0.0289)	0.0612 (0.0491)	0.38
ETR1	0.0342 (0.0272)	0.0364 (0.0200)	0.28
F3H	0.0408 (0.0270)	0.0377 (0.0246)	0.072
FAH1	0.1005 (0.0505)	0.0513 (0.0339)	0.96
HAT4	0.001 (0)	0.0552 (0.063)	3.40
matK	0.356 (0.280)	0.506 (0.317)	1.65
PGIC	0.0556 (0.0254)	0.0393 (0.0235)	3.16
RPM1	0.2096 (0.0235)	0.1678 (0.0476)	0.28

[1], Two times the difference in log likelihoods between a model with fixed *Ka/Ks* ratios among branches, compared with a model with free ratios.
*, p<0.05

2.3.2 Evolution of codon usage bias

Pairwise differences between the species in the presence of major vs. nonmajor codons, as well as overall GC_3 are shown in Table 2.4. In total, there are 32 more cases of *A. lyrata* having a major codon when *A. thaliana* has a nonmajor codon than the reciprocal case, but out of the total of 392 codons which differ, this is not significant (Wilcoxon ranked sign test $p>0.05$). There is also no significant difference in GC_3 between the two species (Wilcoxon ranked sign test, $p>0.05$). Similarly, numbers of unpreferred relative to preferred substitutions do not differ significantly between species using either the total numbers of preferred and unpreferred substitutions ($G=0.02$, $p>0.05$), or considering the subset of loci with sequences from the least diverged outgroups ($G=0.14$, $p>0.05$, $N=5$)

If codon bias is at equilibrium with respect to mutation and selection, an equal number of unpreferred and preferred codons is expected within each lineage. Conversely, if there has been a recent change in mutational biases or selective pressure, the numbers of preferred and unpreferred substitutions will be different (AKASHI 1995; AKASHI 1996). Both species show a similar indication of a slight excess of unpreferred over preferred substitutions; this difference is marginally significant for *A. thaliana* (Tajima's test $\chi^2=4.4$, $p<0.05$), but not for *A. lyrata* ($\chi^2=3.47$, $p>0.05$). When considering the loci sequenced in close outgroup species, however, the difference is significant for both species (*A. thaliana,* 16 preferred, 33 unpreferred, $\chi^2=5.9$, $p<0.05$; *A. lyrata*, 11 preferred, 27 unpreferred, $\chi^2=6.74$, $p<0.05$).

Given the variation in codon bias across loci, some genes may be under weak or no selection on codon usage, while stronger purifying selection may be acting on

others. It is thus worthwhile to analyse differences in codon bias among the species across different categories of overall levels of codon bias. Dividing the genes into high-bias (GC_3>0.4 for both species) and low-bias (GC_3< 0.4 for both species), a significantly higher number of major codon occurrences is observed in *A. lyrata* for low-biased genes (*A. lyrata* total, 90; *A. thaliana* total, 59; Wilcoxon ranked sign test Z=-2.375, p<0.05), but there is no significant difference for high-biased genes (*A. lyrata* total, 112; *A. thaliana* total, 120; p>0.05). Both species show a marginally significant correlation between GC_3 and maximum EST matches, and this correlation is stronger in *A. lyrata*

Table 2.4. Patterns of codon usage bias in *A. thaliana* and *A. lyrata*

locus	$Major_{thal}$	$Major_{lyr}$	GC_3			preferred		unpreferred	
	Non'_{lyr} [1]	Non'_{thal} [1]	*thaliana*	*lyrata*	*outgroup*	*thal*	*lyr*	*thal*	*lyr*
ABAp1	2	2	0.390	0.379	-	-	-	-	-
ABC1At	5	8	0.381	0.393	-	-	-	-	-
ADH	14	11	0.464	0.434	0.449	7	4	6	7
AOP1	6	14	0.328	0.379	-	-	-	-	-
AOP2	6	9	0.317	0.305	0.336	3	0	3	1
AOP3	5	8	0.318	0.331	-	-	-	-	-
AP1	6	4	0.496	0.496	0.513	2	2	2	4
AP3	4	3	0.498	0.481	0.498	2	3	0	2
CAUL	6	4	0.527	0.533	0.492	3	2	1	3
CHI	7	10	0.528	0.564	0.504	4	3	4	2
CHIA	7	11	0.426	0.445	0.462	3	2	9	3
CHS	16	13	0.597	0.627	0.646	4	3	10	11
EnCoH1	6	5	0.324	0.309	0.355	2	1	4	4
ETR1	7	19	0.382	0.389	0.467	4	8	4	4
F3H	18	14	0.518	0.509	0.610	6	5	6	9
FAH1	14	8	0.499	0.483	0.472	6	5	4	8
FKA1	2	8	0.394	0.433	-	-	-	-	-
GL1	2	1	0.401	0.401	-	-	-	-	-
HAT4	2	5	0.415	0.407	0.459	0	1	4	2
matK	-	-	0.254	0.256	0.252	-	-	-	-
PgiC	16	20	0.382	0.397	0.353	4	9	11	5
PIST	5	7	0.490	0.505	-	-	-	-	-
RPM1	19	22	0.403	0.411	0.468	4	5	10	9
Sc-ADH	5	6	0.364	0.376	-	-	-	-	-
total	180	212				54	53	78	74

[1], homologous codons where a major codon is present in one species, and a nonmajor codon in the other. thal, *A. thaliana*; lyr, *A. lyrata*

than *A. thaliana* (*A. thaliana*, Spearman r= 0.402, 1-tailed p<0.05, *A. lyrata*, Spearman r=0.484, 1-tailed p<0.05). However, this correlation is largely due to the *CHS* gene, which has both an exceptionally high GC_3 and number of EST matches (Tables 2.1, 2.2). Excluding this locus, the correlation becomes nonsignificant for *A. thaliana* (r=0.316, p>0.05), and marginally significant for *A. lyrata* (r=0.409, p<0.05). No significant correlation is observed between K_s and the GC_3 values of either species (*A, thaliana* r=0.046, p>0.05; *A. lyrata* r=0.13, p>0.05), suggesting that codon usage is not an important determinant of synonymous substitution rates in this sample of genes.

2.3.3 Evolution of intron size

There is evidence for substantial intron size evolution between *A. thaliana* and *A. lyrata*. Noncoding regions have accumulated a large number of insertion-deletion differences between species, and the cumulative result is a 5% reduction in the amount of DNA derived from intron sequence in *A. thaliana* compared to *A. lyrata* in the 19 genes sampled (16, 883 base pairs in *A. thaliana*, 17,846 in *A. lyrata*). Figure 1 shows the difference in intron size at each locus between the species. Comparing all 87 introns in the dataset, intron size is consistently smaller in *A. thaliana* (Wilcoxon ranked sign test Z=-2.864, p<0.01). Comparing total intron size per gene, however, the difference is not significant in this sample (N=19, Z=-1.932, p=0.053). This difference between total length per gene versus lengths of individual introns probably reflects the fact that several regions sequenced have only a few small introns that show little difference between species (Figure 1), rather than suggesting that the effect is restricted to a small number of loci. Analysis of

alignments suggests that the intron size difference primarily reflects differences in accumulation of small insertions and deletions and simple sequence repeats; no insertion/deletion event between the species from this sample appeared to be the result of a large insertion, such as a transposable element. Since many of the outgroup gene sequences are from cDNA, there is little information on intron size in these species, and only a sample of six genes could be analysed. There is no significant difference in intron size between either species and the outgroup sequence (Wilcoxon signed rank test, $p>0.05$), but the total intron size from this sample of genes was larger than in either *Arabidopsis* species (3287 b.p., compared with 3257 b.p. in *A. lyrata*, and 3091 b.p. in *A. thaliana*).

Estimate of the genomic deleterious mutation rate

Using our pooled estimates of synonymous substitution rates, the divergence time between *A. thaliana* and *A. lyrata* is estimated as between 4.2-10.9 million years ago, using the estimates of substitution rate of 1.5×10^{-8} and 5.8×10^{-9}, respectively. The combined data generate a per site estimate of genomic deleterious mutations between the two species of 0.077. Using these divergence time estimates, U is estimated to be between 0.22 and 0.58.

2.4 Discussion

2.4.1 Effects of mating system on molecular evolution

From the sample of genes examined in this study, there is no evidence for an elevated rate of replacement substitution relative to synonymous substitution in *A. thaliana*, in comparison with its outcrossing congener *A. lyrata*. Similarly, overall levels of codon usage bias do not differ substantially between the species. Both of

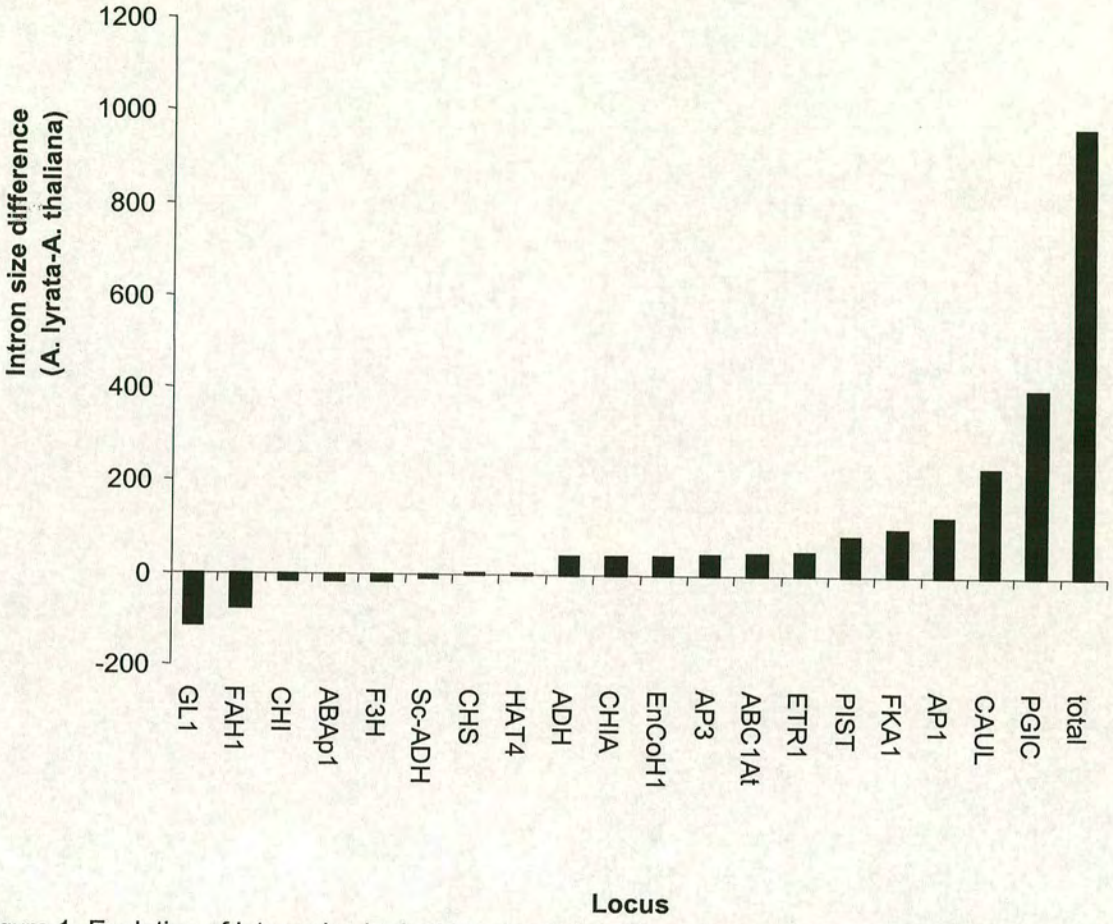


Figure 1- Evolution of intron size in *Arabidopsis*. Total difference in intron size between *A. thaliana* and *A. lyrata* is shown for each locus.

these results conflict with the theoretical prediction of a reduced efficacy of natural selection in self-fertilizing plants. The lack of evidence for an effect of inbreeding is particularly surprising given our high estimate of U, which should generate high levels of background selection in a selfing population (CHARLESWORTH *et al.* 1993). Although the uncertainty of the divergence time means that this mutation rate estimate should be treated with caution, it is within the range of the estimates in *A. thaliana* based on inbreeding depression (CHARLESWORTH 1990), and a mutation accumulation experiment (SCHULTZ *et al.* 1999). In what follows, we discuss several possible explanations for our results, and suggest methods to distinguish between these possibilities.

i) Power to detect substitution differences

One explanation for the similar levels of selective constraint in both lineages is simply a lack of power to detect significant differences among lineages. Our analyses are currently restricted to a small fraction of the genome, and the detection of an effect of breeding system may require substantially more sequence information. However, given that studies using similar samples of genes in other systems have detected significant lineage effects with apparently small differences in effective population size (e.g. Akashi 1996), any effects of breeding system in *Arabidopsis* must be very weak, which is surprising given the potential for large differences in the efficacy of natural selection between self- and cross-fertilizing populations.

Another possibility is that the species used are too divergent to accurately infer substitution rates, and the signal of significant differences in substitution rates

is not detected. Analysis of the subset of genes from the closest outgroup species does suggest a lineage effect on synonymous substitution rates, but does not change our conclusions that selective constraint on amino acid substitutions are the same in the two lineages. Furthermore, all species divergences are well below saturation at replacement sites, and we find no evidence for significant differences in the numbers of amino acid substitutions between species (Table 2.2). Provided that the mutation rate per unit time does not differ greatly between the species, this suggests that there is no major difference in the rate of fixation of amino acids. The conclusions remain unchanged when the maximum likelihood method is used, although approximate estimates of the standard errors of K_a/ K_s were often quite large. In the case of codon bias, our pairwise comparisons of codon usage for a larger sample of genes generated similar conclusions to those based on parsimony, again suggesting that our conclusions are robust for this set of loci.

ii) Nearly neutral vs. neutral models

The lack of effect of breeding system on amino acid substitution and codon bias could be accounted for if there is not a large class of slightly deleterious mutations in *Arabidopsis*, so that most synonymous and amino acid substitutions that have fixed since the divergence of the species studied were effectively neutral with respect to fitness in both species. Alternatively, mutations with small deleterious effects in heterozygotes may occur, but they may be nearly recessive, and experience strong selection in homozygotes, preventing their spread in selfing populations. Although the reduction in effective size due to background selection is thought to outweigh the purging effects of high levels of homozygosity

(CHARLESWORTH 1994), theoretical investigation of the interaction of these effects remains preliminary, and the quantitative importance for the fixation of deleterious mutations in selfers versus outcrossers will depend on the distributions of selective effects and dominance coefficients. The strength of background selection also depends on the deleterious mutation rate, so another possibility is that this may be too low to substantially reduce N_e. However, our high estimate of U based on sequence data, and comparable estimates using levels of inbreeding depression, make this explanation seem unlikely.

Conclusions based on a 'random' sample of loci are also complicated by the presence of differences in the strength and direction of selection among genes and individual sites. Our analysis relied on pooling a heterogeneous set of loci, which clearly vary in their levels of selective constraint, and some of these loci may also have been subject to positive selection on amino acid mutations. Because of the substantial differences among these loci, the effects of breeding system and recombination rate are very likely to have differential effects on these different classes of genes. In the case of codon bias, we observed higher levels of major codon usage in *A. lyrata* when only the low-biased genes are examined. This may reflect strong selection ($N_e s >> 1$) on highly biased genes in both species, whereas low-biased genes are probably under weaker selection, allowing more deleterious mutations to accumulate in selfing lineages. The effects of variation in selection coefficients should be investigated in more detail by sampling a larger number of weakly and highly expressed genes in *A. lyrata* and an outgroup species, and comparing codon bias and amino acid substitution in the two lineages for these different classes of loci.

iii) Population size changes

Models predicting an accumulation of slightly deleterious mutations in selfing populations assume that population size has remained constant in both species. However, if *A. lyrata* has undergone a population bottleneck, rates of amino acid substitution and unpreferred codon substitution could have been reduced, obscuring any differences when compared with the inbreeding species *A. thaliana*. Such a population bottleneck could have been associated with post-glacial recolonization of northern Europe and North America (COMES and KADEREIT 1998). Consistent with a bottleneck hypothesis, our preliminary evidence suggests a departure from equilibrium codon usage in both species, with an excess of unpreferred substitutions. This suggests a reduction in codon usage bias in both lineages since divergence from their ancestor, similar to recent conclusions from comparisons of codon usage in *D. melanogaster* and *D. simulans* (MCVEAN and VIEIRA 2001). However, a shift in mutational bias or codon preference in *A. lyrata* remains a possibility.

Studies of polymorphism at the *ADH* locus in *A. lyrata* (SAVOLAINEN *et al.* 2000) found an excess of intermediate frequency variants in North American populations, which is consistent with a bottleneck hypothesis, but this was not evident from smaller samples of European populations at this locus, although sample sizes were small. Our own data on polymorphism in European populations at several other loci, has also not found any such result (see Chapter 4). The possibility that effective population size has been reduced in both species can be tested by gathering more data on polymorphism in *A. lyrata*, and by examining a larger sample of outgroup species, to test for elevated substitution rates in both species.

Conversely, if absolute population size has increased in *A. thaliana* since the evolution of selfing, the efficacy of selection may not be low. Species-wide samples of polymorphism in *A. thaliana* indicate a frequent skew in the frequency spectrum towards rare variants, a unimodal distribution of pairwise differences (KUITTINEN and AGUADE 2000), and a lack of genetic isolation by distance (BERGELSON *et al.* 1998; but see SHARBEL *et al.* 2000), as expected under a recent population expansion. However, none of these features is consistent across all loci, and it is unclear to what degree the frequency of rare variants simply reflects the presence of strong population structure rather than global population size changes. It is also likely that such recent historical expansion events occurred too recently to affect patterns of molecular evolution, and a historical signature of higher rates of slightly deleterious fixation would therefore still be expected.

A problem with explanations based on population size change is the evidence for highly reduced levels of polymorphism within populations of *A. thaliana*, and the evidence of strongly subdivided populations (ABBOTT 1989; BERGE *et al.* 1998; BERGELSON *et al.* 1998). The effects of breeding system on the efficacy of selection in strongly subdivided populations remain poorly understood, and it is unclear what forms of migration and selection would allow selection to be effective species-wide in *A. thaliana*. Given evidence for similar levels of species-wide polymorphism in both inbreeding and outbreeding *Arabidopsis* species (SAVOLAINEN *et al.* 2000), it is possible that strong population structure allows locally high frequencies of deleterious mutations, while preventing fixation species-wide. One might then frequently find an excess of replacement polymorphism in species-wide samples, as observed in *A. thaliana* (KAWABE *et al.* 1997; PURUGGANAN and SUDDITH 1998;

PURUGGANAN and SUDDITH 1999), in contrast to the pattern in Drosophila nuclear genes (WEINREICH and RAND 2000). However, the effect of high levels of homozygosity on the purging of deleterious mutations from local populations is uncertain. Clearly, more theoretical investigations of these effects, and more detailed analyses of polymorphism and population subdivision in both species, are necessary to evaluate the interaction between breeding system and population structure in influencing the efficacy of natural selection.

iv) How long has A. thaliana been self-fertilizing?

It is often suggested that selfing species persist for only short evolutionary times (reviewed in TAKEBAYASHI and MORRELL 2001), and rapidly go extinct. If most selfing lineages are very recently derived, it may be difficult to detect deleterious mutation accumulation. However, if self-fertilizing populations become extinct before substantial mutation accumulation has occurred, the genetic explanation for their short evolutionary life spans (TAKEBAYASHI and MORRELL 2001) cannot be correct. The amount of time during which *A. thaliana* has been self-fertilizing is unknown, but as a maximum estimate, the divergence between *A. thaliana* and its close relatives has been estimated as between 3.1 and 9 million years ago (KOCH *et al.* 2000). However, it is much more difficult to assess the minimum time during which the species has been self-fertilizing. Comparisons of the putative *A. lyrata* self-incompatibility locus (*SRK*) indicate that the most similar *A. thaliana* sequence encodes a truncated kinase domain, and the locus is probably a pseudogene (KUSABA *et al.* 2001). Such mutations at this locus may have caused *A. thaliana*'s self-fertility, or could have occurred after self-fertilization evolved.

However, it is impossible to infer the precise date of this event, due to the high silent and replacement polymorphism among *A. lyrata* alleles (SCHIERUP *et al.* 2001). Nevertheless, high silent and amino acid divergence between species using the most similar *A. lyrata SRK* allele (C. Bartholomé and D. Charlesworth, unpublished data) suggest that this locus may have been nonfunctional for a long time, and therefore that *A. thaliana* did not become self-fertile long after its separation from *A. lyrata*. As further information becomes known about genome-wide patterns of linkage disequilibrium in *A. thaliana* (NORDBORG *et al.* 2002), the data might be used to estimate the historical rate of self-fertilization (NORDBORG 2000), although this is also complicated by the presence of population structure.

2.4.2 Effects of Gene Expression on Patterns on Molecular Evolution

A surprising result from our study is that gene expression level explains a large proportion of the observed variance in selective constraint on amino acids among loci. Using the number of EST matches as a crude estimate of gene expression, we observe a strong correlation with amounts of amino acid, but not silent, substitutions, even for loci with a complete cDNA sequence. This suggests that less expressed genes either have low selective constraints on amino acid substitutions, or else that they are more likely to be subject to positive selection in *Arabidopsis*. In mammals, there is a correlation between the breadth of expression and K_a/K_s, which has been interpreted as evidence that a higher proportion of replacement changes affect function in genes expressed in many tissues (DURET and MOUCHIROUD 2000). While the EST database for *A. thaliana* does not include sufficient sampling across tissues to investigate this in detail, breadth of expression

may generally be associated with the overall level of gene expression (AKASHI 2001). As more quantitative information on gene expression becomes available, it will be important to investigate in more detail the effects of expression levels on substitution rates and gene structure.

2.4.3 Evolution of intron size in Arabidopsis

In our sample of loci, intron size was consistently smaller in *A. thaliana* than *A. lyrata*. This contrasts with patterns observed in *Drosophila melanogaster*, where regions of low recombination have larger introns on average (CARVALHO and CLARK 1999; COMERON and KREITMAN 2000). However, it is consistent with measurements of genome size in the two study species; *A. lyrata* is estimated to have a larger genome size than *A. thaliana* (O. Savolainen, pers. comm.). Comparative mapping of a BAC (bacterial artificial chromosome) clone containing the putative self-incompatibility locus in *A. lyrata* has also provided evidence that intergenic sequences are consistently larger in comparison with those of *A. thaliana* (KUSABA *et al.* 2001), so a general decrease in the sizes of noncoding regions is possible in *A. thaliana*. The contrast may reflect directional selection on intron size in a fast-growing annual, given the observed negative correlation in plants between genome size and weediness (BENNETT 1998). However, evidence for high rates of deletion in the *ABC1At* pseudogene in *A. lyrata* indicates that mutation may rapidly eliminate 'junk' DNA, suggesting that selection may be maintaining large intron size in these species. Comparisons of segregating insertions and deletions with those that are fixed between species should be helpful in distinguishing between effects of mutational biases and selection (COMERON and KREITMAN 2000), although the

estimation of numbers of insertion and deletion events is a challenge even for modestly diverged species.

Chapter 3

Testing for selection on synonymous and replacement mutations in *Arabidopsis thaliana* and *Arabidopsis lyrata*

3.1 Introduction

In a previous study of 23 genes, we compared levels of codon usage bias and rates of nucleotide substitution between the outcrossing *Arabidopsis lyrata* and the self-fertilizing *A. thaliana* (Chapter 2). Although we observed a trend towards more preferred codons in *A. lyrata* the effect was nonsignificant, and we found no evidence for a higher rate of preferred relative to unpreferred changes, or a lower rate of replacement relative to synonymous substitution in *A. lyrata*. We thus found little evidence for the difference in the effectiveness of selection between selfing and outcrossing species expected under a nearly neutral model of molecular evolution. However, the power of our tests may have been limited for several reasons. First, the inherent noise in the process of mutation-selection-drift may make it difficult to detect a significant effect in small numbers of loci. Similarly, variation across loci in levels of selective constraint could mask the signature of differences in selection. Thirdly, with this sample size it was not possible to investigate codon preferences in *A. lyrata*, in order to test the assumption that they are similar to those in *A. thaliana*.

In addition to the question of statistical power, rates of substitution in this study were estimated using a single sequence for each species. Some of these identified changes will thus represent new mutations segregating within species,

which have not yet been subject to the action of natural selection. For example, under a mutation-selection-drift model for codon bias, an excess of unpreferred over preferred polymorphism is expected at equilibrium, because of the high proportion of preferred codons within genes. Natural selection will then act to remove unpreferred mutations from the population, leading to an equal ratio of preferred relative to unpreferred fixations (AKASHI 1997). When a single sequence from each species is used to estimate substitution rates, the rate of unpreferred substitutions may then be overestimated, due to the contribution of segregating mutations. This effect could contribute to the observed excess of unpreferred substitutions in both *Arabidopsis* species (Chapter 2). Furthermore, a consistent excess of amino acid polymorphism relative to fixed differences has been observed in *A. thaliana* (WEINREICH and RAND 2000). Combined with our results, this suggests that while the product of the species-wide effective size and the selection coefficient may be high enough to prevent a high rate of fixation of amino acids, the high selfing rate in *A. thaliana* may allow many replacement changes to segregate in natural populations. This suggests the importance of incorporating polymorphism into the comparison of both replacement and synonymous mutations in *Arabidopsis*.

Although a positive correlation between preferred codon frequencies and gene expression has been detected for both species (Chapter 2; DURET and MOUCHIROUD 1999), the corresponding correlation coefficients are weak, and codon preferences have been defined solely using genome-wide multivariate analysis of codon usage, by identifying codon preferences in genes at the extreme of the first axis of correspondence analysis. Thus, there remains little independent evidence in either *Arabidopsis* species that the identified 'preferred' codons are explained by

translational selection. While a strong association between tRNA abundance and codon bias has been repeatedly demonstrated in bacteria and yeast (IKEMURA 1985), there is less evidence for such an association in multicellular eukaryotes. In Drosophila, there appears to be a general correspondence between experimentally estimated tRNA abundance and codon bias (MORIYAMA and POWELL 1997). Similarly, in *C. elegans*, when tRNA gene copy number is used as a proxy for tRNA abundance, there is a good concordance between the most frequently used codons and the copy number of the corresponding issoaccepting tRNAs (DURET 2000), although there has been subsequent debate over the codon-tRNA pairing assumptions used in the analysis (PERCUDANI 2001).

In this study, I reinvestigate the effectiveness of natural selection at replacement and synonymous sites in *A. lyrata* and *A. thaliana*, using a larger sample of loci (83 genes), and incorporating a comparison of polymorphism and divergence at synonymous and replacement sites between species. I also take advantage of new fine-scale estimates of gene expression in *A. thaliana* using Massively Parallel Signature Sequencing (MPSS; BRENNER *et al.* 2000), and data on tRNA gene abundance in the *A. thaliana* genome (http://rna.wustl.edu/GtRDB/At/). I use these data to address four primary questions: 1) Does translational selection act on codon bias in *Arabidopsis*? 2) Are codon preferences conserved between species? 3) Is there higher codon bias and a higher rate of fixation of preferred codons in the outcrossing *A. lyrata*? 4) Is there an excess of amino acid polymorphism relative to divergence in *A. thaliana* compared with *A. lyrata*?

3.2 Methods

Datasets

The complete list of 83 loci used for the analysis of codon bias is shown in Appendix A.1. We include all genes from the original study, with the exception of two members of the AOP gene family (AOP1 and AOP2), to avoid redundancies of base composition and substitution patterns across gene family members. We have added 22 new *A. lyrata* genes that are now available from Genbank, eight genes provided by colleagues (E. Kamau, H. Kuittinen and O. Savolainen, pers. comm.), as well as 32 new genes sequenced for this study. For comparisons of polymorphism and divergence, we use several polymorphism datasets for each species, as indicated in Appendix A.1. First, we use polymorphism data from 5 nuclear genes sequenced in species-wide samples of *A. lyrata* and *A. thaliana* (Chapter 4). In addition, partial fragments of these five genes have been sequenced (H. Zheng and M. Nordborg pers. comm.) in a much more extensive worldwide sample of *A. thaliana*, and we include these data in the analysis of replacement and synonymous polymorphism. We also make use of the published polymorphism datasets available from *A. thaliana* (see Chapter 4), and our direct sequencing survey of coding sequence polymorphism for 18 exons in an Icelandic population of *Arabidopsis lyrata* (Chapter 6). To obtain outgroup sequence for parsimony analyses, genes were submitted to a BLAST search against the *Brassica oleracea* shotgun genome sequencing project. Sequences were aligned using Genedoc, and for a subset of loci we used these alignments to identify conserved regions for PCR amplification in the genus *Capsella* (*C. rubella* or *C. grandiflora*), to get a close outgroup for parsimony

analysis. For all other loci, and when PCR or sequencing failed in *Capsella*, we use *Brassica* sequence for outgroup information.

Data Analysis

Calculations of polymorphism and divergence for preferred, unpreferred, synonymous and replacement mutations were made following strict parsimony assumptions, using DnaSP 3.95 (ROZAS and ROZAS 1999). To investigate codon preferences, we conduct correspondence analysis of relative synonymous codon usage (RSCU) of the sequenced *A. lyrata* genes, using the program CodonW (http://www.molbiol.ox.ac.uk/cu/culong.html). We exclude all gene fragments with less than 100 synonymous sites for this analysis, since small sequences can generate random biases in codon usage. Following the approach of Chiapello *et al.* (1998) for *A. thaliana*, preferred codons were detected by comparing codon frequencies of the highest- and lowest- biased genes, defined by the two extremes of the first axis of the correspondence analysis. Chi squared analysis was performed to test the null hypothesis that the relative frequencies of individual codons are the same in the highest- and lowest-biased 30 genes, against the alternative hypothesis of an increase in frequency. We compare these results to the genome-wide identification of codon preferences in *A. thaliana*. For further comparison, we also performed correspondence analysis in *A. thaliana* on the orthologous regions we examined in *A. lyrata*, to evaluate the assessment of codon preferences in this subset of genes. CodonW was also used to estimate F_{op}, the frequency of optimal codons (STENICO *et al.* 1994) for each gene.

Data on tRNA gene abundance in the complete genome of *A. thaliana* was obtained from the genomic tRNA database (http://rna.wustl.edu/tRNAdb/). To compare with preferred codon frequencies, we group codons by their predicted major tRNA species, following the revised wobble rules for eukaryotic genomes (see PERCUDANI 2001). These rules assume that GNN tRNAs pair with both C- and U- ending codons, while ANN tRNA genes are modified to inosine, and decode both U- and G- ending codons. Identified tRNA genes in *Arabidopsis thaliana* generally adhere to these wobble rules, since almost every codon can only be decoded by a single class of tRNA (see Table 3.1). However, there are three exceptions of single unexpected tRNA genes (shown in italics in Table 3.1- Gly GGU, Ser GGA, Leu GAG), which have been hypothesized to represent pseudogenes or sequencing errors (WANG *et al.* 2002). For our estimates of relative tRNA abundance, we ignore these exceptional genes, although their inclusion does not change the rank order of tRNA gene frequencies. tRNA gene abundance has been found to correlate strongly with the corresponding tRNA abundance in a number of prokaryotic and eukaryotic genomes (KANAYA *et al.* 2001), suggesting that it is a good proxy for relative tRNA abundance in the genome. In accordance with this prediction, we find a moderately significant positive correlation between codon number in the high-biased genes from Chiapello *et al.* (1998), and the number of isoaccepting tRNAs (r=0.40, p<0.01). When we remove three outliers with the highest tRNA gene copy numbers, the correlation improves substantially (r=0.84, p<<0.001).

We use measures of gene expression level for each gene from *A. thaliana* using two datasets. First, data on the maximum abundance of expressed sequence tags (ESTs) were obtained from G. Marais. Second, quantitative estimates of gene

expression from 5 tissues were obtained from a database of Massively Parallel Signature Sequencing (BRENNER *et al.* 2000) of mRNA in *A. thaliana* (http://dbixs001.dbi.udel.edu/MPSS4/java.html). This method generates quantitative estimates of the abundance of short sequence tags from messenger RNA. To estimate expression level for each gene, we searched the *Arabidopsis* MPSS database in two ways. First, the search by protein identifies sequence tag matches in the vicinity of the queried gene. We estimated expression level for each tissue by summing all expressed tags in the sense strand either within the coding sequence of the annotated gene (Class I) or within 500bp 3' of the coding region (Class II). However, because this database searches for sequence tag matches to genomic DNA, it does not include possible tags generated after intron splicing. We therefore also submitted the available full-length mRNA for each locus to a search in the database to identify additional expressed sequence tags, and the abundance of any additional tags were added to estimates of expression level. For a small subset of loci, expressed tags matched to multiple genes, and these loci were therefore excluded from subsequent analysis. To investigate the effects of gene expression, we use both the maximum expression level across tissues and the total expression level across tissues as our estimates. For genome-wide analysis of the *A. thaliana* genome, gene-by-gene estimates of MPSS expression level were provided directly by B. Meyers. These data only represent expression levels for sequence tags matching the genomic DNA, and might therefore in some cases represent underestimates of the gene expression level.

3.3 Results and Discussion

3.3.1 Codon preferences in Arabidopsis lyrata

A summary of inferred codon preferences in *A. lyrata* is shown in Table 3.1, with the *A. thaliana* results for the orthologous genes and the genome-wide results of Chiapello *et al.* (1998) given for comparison. In comparison with the results of Chiapello *et al.* (1998) we identify a moderate concordance between identified codon preferences; for 13 out of 21 'preferred' codons in *A. thaliana* we identify the same codons as preferred, while three additional codons show a nonsignificant increase in frequency. Conversely, a single codon that is unpreferred in *A. thaliana* was identified as preferred in *A. lyrata* (ACG). However, comparison with correspondence analysis from *A. thaliana* using the same set of loci shows the identification of very similar codon preferences to those in *A. lyrata*; the rank order of relative codon frequencies is identical across the two species, and relative codon frequencies are very highly correlated (Spearman's r=0.96, p<<0.01). This suggests that differences from the genome-wide analysis may reflect artifacts of multivariate analysis on a smaller subset of genes, rather than true differences in codon preferences between species. Thus, this dataset gives little evidence for changes in codon preferences between species.

3.3.2 Evidence for translational selection favouring preferred codons

To test the hypothesis that the identified preferred codons are under translational selection, we investigated the correspondence between codon preferences and tRNA gene abundance in the *A. thaliana* genome. tRNA gene numbers, relative codon frequencies between high- and low-biased genes, and

absolute frequencies in highly biased genes are shown in Table 3.1. By combining codons using the same tRNA isoacceptor, we find a perfect correspondence between the most abundant tRNA gene class and codon classes showing the highest relative frequencies in high-biased genes.

Table 3.1 Codon preferences and tRNA gene abundance in *Arabidopsis*.

		# of tRNA genes	relative codon freq. lyrata[1]	relative codon freq. thaliana[1]	relative codon freq. *thaliana*[2] (whole genome)	relative codon freq. by isoacceptor[3]	tRNA relative abundance by isoacceptor[4]
Gly	GGA	12	0.99	1.02	0.979 (0.382)	0.980 (0.382)	0.876
	GGC	23	**2.58**	**1.97**	**1.49 (0.158)**	**1.23 (0.561)**	**1.75**
	GGU	*1*	0.77	0.85	***1.15*** **(0.403)**		
	GGG	5	0.81	0.77	0.361 (0.056)	0.36 (0.056)	0.37
Val	**GUU**	15	0.8	0.84	0.86 (0.377)	**1.26 (0.722)**	**1.5**
	GUC		**1.96**	**1.78**	**2.54 (0.345)**		
	GUA	7	0.76	0.81	0.29 (0.050)	0.29 (0.050)	0.7
	GUG	8	0.98	0.92	0.90 (0.228)	0.90 (0.228)	0.8
Lys	AAA	13	1.03	1.0	0.42 (0.217)	0.42 (0.217)	0.8
	AAG	18	0.97	1.0	***1.6*** **(0.783)**	**1.6 (0.783)**	**1.2**
Asn	AAT		0.68	0.68	0.310 (0.185)	-	-
	AAC	16	**1.45**	**1.45**	**2.02 (0.815)**		
Gln	CAA	8	1.06	1.13	0.692 (0.393)	0.692 (0.393)	0.9
	CAG	9	0.94	0.85	***1.41*** **(0.607)**	**1.41 (0.607)**	**1.1**
His	CAC	10	**1.95**	**1.76**	**3.10 (0.749)**	-	-
	CAT		0.61	0.71	0.331 (0.251)		
Glu	GAA	12	1.02	1.01	0.64 (0.336)	0.64 (0.336)	0.96
	GAG	13	0.98	0.99	***1.40*** **(0.664)**	**1.40 (0.664)**	**1.04**
Asp	GAC	23	**1.76**	**1.75**	**1.97 (0.535)**	-	-
	GAT		0.75	0.73	0.637 (0.465)		
Tyr	*TAC*	76	**1.7**	**1.53**	**2.41 (0.862)**	-	-
	TAT		0.61	0.65	0.21 (0.138)		
Cys	TGC	**15**	**1.92**	**1.52**	**1.83 (0.626)**	-	-
	TGT		0.71	0.79	0.25 (0.374)		
Phe	*TTC*	16	**1.3**	**1.48**	**2.06 (0.740)**	-	-
	TTT		0.75	0.67	0.41 (0.260)		
Ile	ATA	5	0.87	0.87	0.21 (0.061)	0.21 (0.061)	0.4
	ATT	20	0.69	0.69	0.80 (0.361)	**1.32 (0.939)**	**1.6**
	ATC		**1.67**	**1.68**	**2.26 (0.578)**		
Arg	*AGA*	9	0.96	1.04	0.70 (0.261)	0.70 (0.261)	**1.3**
	AGG	8	0.57	0.57	***1.18*** **(0.248)**	1.18 (0.248)	1.1
	CGA	6	1.25	0.89	0.41 (0.050)	0.41 (0.050)	0.83
	CGG	4	2.11	2.25	0.29 (0.030)	0.29 (0.030)	0.56
	CGU	9	1.15	1.23	**2.22 (0.313)**	**2.18 (0.411)**	**1.3**
	CGC		1.44	1.18	**1.67 (0.098)**		
Leu	CTA	10	1.07	1.23	0.64 (0.036)	0.64 (0.036)	1.2
	CTG	3	1.14	1.09	0.47 (0.061)	0.47 (0.061)	0.37
	CTT	11	0.76	0.81	1.04 (0.284)	**1.5 (0.592)**	**1.46**
	CTC	*1*	**2.23**	**2.27**	**2.59 (0.308)**		

	TTA	6	0.74	0.72	0.25 (0.036)	0.25 (0.036)	0.73
	TTG	10	0.80	0.74	1.06 (0.242)	1.06 (0.242)	1.2
Ser	AGC	13	1.08	1.26	1.37 (0.171)	0.79 (0.270)	0.812
	AGT		0.65	0.64	0.55 (0.099)		
	TCA	9	0.88	0.89	0.73 (0.154)	0.73 (0.154)	0.56
	TCG	4	1.27	0.92	0.72 (0.070)	0.72 (0.070)	0.25
	TCT	37	1.06	1.16	0.93 (0.275)	**1.31 (0.507)**	**2.38**
	TCC	*1*	1.42	1.48	**2.52 (0.232)**		
Thr	ACA	8	0.74	0.73	0.55 (0.175)	0.55 (0.175)	1.0
	ACG	6	1.59	1.38	0.49 (0.067)	0.49 (0.067)	0.75
	ACT	10	0.78	0.78	0.90 (0.335)	**1.45 (0.758)**	**1.25**
	ACC		1.51	**1.88**	**2.86 (0.423)**		
Pro	*CCA*	39	0.98	0.94	1.10 (0.375)	**1.10 (0.375)**	**1.95**
	CCG	5	1.18	1.51	0.60 (0.113)	0.60 (0.113)	0.3
	CCT	16	0.74	0.71	0.80 (0.307)	1.08 (0.511)	0.8
	CCC		**1.82**	**1.70**	**2.32 (0.204)**		
Ala	GCA	10	0.74	0.70	0.50 (0.142)	0.5 (0.142)	0.91
	GCG	7	1.23	1.13	0.64 (0.084)	0.64 (0.084)	0.64
	GCT	16	0.86	0.82	1.05 (0.484)	**1.32 (0.774)**	**1.45**
	GCC		**2.0**	**2.59**	**2.32 (0.290)**		

[1] relative frequency of codons in high- relative to low-biased genes from correspondence analysis of loci with both *A. lyrata* and *A. thaliana* orthologs sequenced. Values in bold represent significant increases in relative frequency in high biased genes.
[2] relative codon frequencies from Chiapello *et al.* (1998). Values in parentheses are the absolute frequencies in high biased genes.
[3] values in bold represent codon classes with the highest relative frequency in high- vs. low-biased genes.
[4] values in bold represent the highest tRNA relative abundance

Furthermore, with the exception of two six-fold degenerate codons (Leu and Arg), the rank order of codon frequencies corresponds exactly with the rank order of tRNA abundance. This result provides strong support for the hypothesis that the preferred codons are favoured by translational selection. Interestingly, four of the five preferred codons that do not show an increase in frequency in our smaller dataset (AAG, CAG, GAG, and AGG) have only slight differences in tRNA abundance. This suggests that, as would be expected a priori, the smaller dataset has less power to detect subtler differences in selection.

For codons recognized by the same tRNA, Ikemura (1985-'rule 4') proposed that because of weak pairing interactions, tRNAs pairing with (A/U)-(A/U)-pyrimidine codons should favour C-ending codons. This rule appears to hold in the *A. thaliana* genome, supporting the hypothesis of preference for the C-ending

codons in Asn, Tyr, Phe, and Ile. However, in other cases (Val, Ser, Thr, Pro, His, Cys), there is an apparent preference of C- over U- ending codons that is not expected according to these rules. Furthermore, there are several other cases where the preferred and unpreferred codons identified by Chiapello *et al.* do not match precisely with the expectations based on tRNA abundance (Arg, Leu, Pro, Ala-see Table 3.1). Because low-biased genes are generally AT-rich, these apparent disagreements could represent artifacts of defining preferred codons as those that show a significant increase in frequency at one extreme of the first multivariate axis. Since the first multivariate axis maximizes the differences between genes in codon usage, it may be influenced by base composition as well as selection. For example, when the absolute frequencies of GCT, GTT, TCT, ACT and CCT in high biased genes are examined, no such preference for C- over U-ending codons is apparent (in contrast with the ratio of usage in high vs. low-biased genes). Thus, it is possible that the identification of preferred and unpreferred codons using multivariate analysis could lead to the misidentification of some codons. Alternatively, there may be an additional selection pressure favouring C- ending codons such as mRNA secondary structure. This possibility has been hypothesized in *Drosophila melanogaster* to explain the strong preference for C-ending codons (CARLINI *et al.* 2001), which is not always expected from information on tRNA gene abundance (KANAYA *et al.* 2001).

To distinguish these hypotheses, we compare two estimates of the frequency of optimal codons: one based on codon preferences identified in Chiapello *et al.* (1998) using multivariate analysis (Fop_{mult}), and a second based on expectations derived from tRNA gene relative abundance (Fop_{tRNA}). For this latter measure, we

identify preferred codon classes as those with the highest relative tRNA abundance for each amino acid. We also follow Ikemura's rule 4 (1985) and assume a preference of C over T for (A/U)-(A/U)-pyrimidine codons. Preferred codons defined by these rules are shown in bold in Table 3.1. Correlation analysis between three measures of gene expression, these two estimates of F_{op}, and the GC content at third codon positions (GC_3) is shown in Table 3.2. First, as expected, Fop_{tRNA} shows a weaker correlation with GC_3 (r=0.51) than Fop_{mult} (r=0.87), due to the higher proportion of A/T-ending preferred codons. For all three measures of codon usage, MPSS estimates of gene expression tend to show a slightly higher correlation with codon bias than EST-based estimates.

For both *A. lyrata* and *A. thaliana*, Fop_{tRNA} shows the highest correlation with all measures of gene expression. Furthermore, a partial correlation between Fop_{mult} and gene expression, factoring out Fop_{tRNA}, is not significant, suggesting that there is no residual effect of this measure of codon bias once the similarity with Fop_{tRNA} is factored out (Table 3.2). Conversely, there is still a residual effect of gene expression on Fop_{tRNA}, factoring out Fop_{mult} (Table 3.2). Finally, GC_3 is not positively correlated with gene expression in a partial correlation, after factoring out the effects of Fop_{tRNA}. Since selection on mRNA secondary structure is likely to be associated with gene expression level, this suggests that there is not a residual action of selection on base composition, independent of translational selection. These results suggest that the codon preferences identified using tRNA relative abundance more accurately reflect those codons favoured by translational selection than those identified using multivariate analysis, and that codon preferences identified solely using correspondence analysis should be treated with caution.

Table 3.3 Pearson correlations between gene expression and codon usage bias[1]

Sample	Codon Bias measure	Maximum expression, EST	Maximum expression, MPSS	Total Expression, MPSS
A. thaliana, whole genome	Fop_{tRNA}	0.2977***	0.2681***	0.3201***
	partial	0.2088***	0.1741***	0.2244***
	Fop_{mult}	0.2171***	0.2085***	0.2345***
	partial	0.0065	0.0250**	0.0085
	GC_3	0.1746***	0.1564***	0.1686***
	partial	-0.0536***	-0.0312***	-0.0759***
A. thaliana, 83 genes	Fop_{tRNA}	0.2676*	0.3406**	0.3491**
	partial	0.2681*	0.2413*	0.2679*
	Fop_{mult}	0.1339	0.2752*	0.2353*
	partial	-0.1096	0.0155	-0.0384
	GC_3	0.1390	0.2304*	0.2043
	partial	0.0047	0.0836	0.0309
A. lyrata, 83 genes	Fop_{tRNA}	0.2567*	0.3376**	0.3341**
	partial	0.2318*	0.2284*	0.2445*
	Fop_{mult}	0.1480	0.2707*	0.2366*
	partial	-0.0783	0.0133	-0.0286
	GC_3	0.1727	0.2641*	0.2386*
	partial	0.0390	0.1112	0.0626

[1]partial correlations control for Fop_{mult} when examining Fop_{tRNA}, and control for Fop_{tRNA} when correlating gene expression with Fop_{mult} and GC_3.
*, $p<0.05$; **,$p<0.01$, ***, $p<0.001$

3.3.3 Comparing codon usage bias in A. lyrata vs. A. thaliana

For pairwise comparisons of codon usage, we compare three different measures of codon bias; a) GC_3, b) the number of orthologous codons where *A. lyrata* (*A. thaliana*) has a preferred codon, and *A. thaliana* (*A. lyrata*) has an unpreferred codon using the multivariate definitions of codon preferences (see chapter 2), and c) the same comparison using tRNA abundance- based definitions of preferred codons. Using all 83 loci, there is a significantly higher GC_3 in *A. lyrata* compared with *A. thaliana* (Figure 1; Wilcoxon ranked sign test Z=-2.395, one-tailed $p<0.001$; 2-tailed $p<0.05$). Similarly, we observe a marginally significant increase in the number of preferred codons in *A. lyrata* using multivariate-based measures of codon preference (Wilcoxon ranked sign Z=-1.636, 1-tailed $p<0.05$, 2-tailed $p=0.1$). In contrast, we find no significant difference between species in a

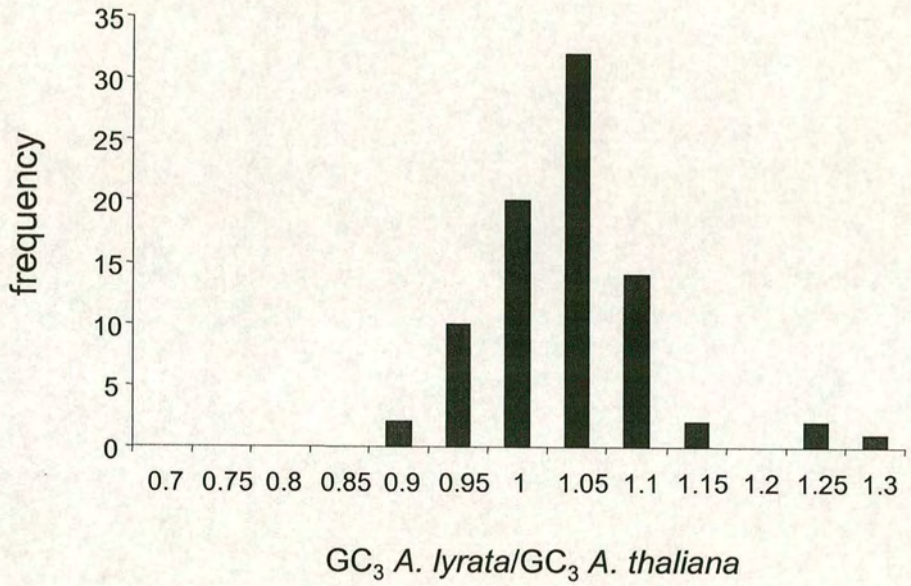


Figure 3.1- Frequency distribution of the ratio of GC content at third codon positions in *A. lyrata* relative to *A. thaliana*.

comparison of tRNA-based codon preferences (Wilcoxon ranked sign Z=-0.4199, p>>0.05), although there is a weak trend towards a higher number of preferred codons in *A. lyrata* (40 genes preferred codons greater in *A. lyrata*, 32 genes in *A. thaliana*). Thus, while we find evidence for a significant change in codon usage between species, this difference is not consistent for all three measures of codon bias. Levels of codon bias in *Arabidopsis* depend both on gene expression level and on gene length (DURET and MOUCHIROUD 1999), and this might influence our comparison across species. In particular, long and lowly expressed genes may be evolving neutrally with respect to codon bias, adding noise to the comparison of selection on codon usage. If we restrict our analysis to genes with less than 500 codons and with evidence for expression from either MPSS or EST data, we observe a more significant difference in both GC_3 (2-tailed p<0.01) and multivariate-based comparisons of codon preference (2-tailed p<0.05) towards higher codon bias in *A. lyrata*, but still no significant difference in tRNA-based estimates of codon preference (p>>0.05; 19 genes preferred codons greater in *A. lyrata*; 12 genes greater in *A. thaliana*).

3.3.4 Selective vs neutral explanations for differences in codon bias

The observed differences in codon bias between species could result from differences in the strength of natural selection, or they could derive from neutral differences among species in rates of substitution. Given that the difference between species is strongest for GC_3, which appears to be the measure least associated with translational selection (Table 3.2), it is possible that the observed differences reflect a generally higher GC content in *Arabidopsis lyrata*, rather than stronger selection

on codon bias. This elevated GC might reflect either a difference in mutation preferences, or a higher rate of biased gene conversion in *A. lyrata*. However, we do not have data on tRNA gene abundances in *A. lyrata*, and it may be that the definitions of codon preferences apply less well to this outcrossing species. Under a neutral model, a similar elevated GC content would be expected in noncoding regions, as was observed in a comparison of *Brassica rapa* with *A. thaliana* (Tiffin and Hahn, 2002). However, we find no significant difference in GC_{intron} between the species for the genes with introns sequenced in *A. lyrata* (Wilcoxon ranked sign test Z=-0.8863, p>>0.05 n=43), and in fact the trend in the data is towards a higher GC_{intron} in *A. thaliana* (9 more genes had greater GC_{intron} in *A. thaliana* than in *A. lyrata*). We thus find no evidence for a general elevation in GC content in *A. lyrata*. However, intron sequences are short in *Arabidopsis*, and it is possible that a significant proportion of intron sequence is under selective constraint in such species (MARAIS and PIGANEAU 2002). Indeed, in the complete *A. thaliana* genome, GC_{intron} is negatively correlated with GC_3 (r=-0.223, p<0.001), suggesting that base composition in introns may be a poor predictor of neutral base composition.

To further distinguish between neutral and selection-based explanations for the difference in codon usage between species, we examine patterns of polymorphism and divergence at synonymous sites. In particular, differences in mutation preference should generate a difference among species in the ratio of GC-AT relative to AT-GC polymorphisms, while stronger selection or a higher rate of biased gene conversion should generate a difference between species in the ratio for fixed differences compared with the ratio for polymorphism. Table 3.3 shows the results of calculations of polymorphic and fixed mutations for the multivariate and

tRNA-based definitions of preferred and unpreferred changes, and for AT-GC vs. GC-AT synonymous changes. Because of differences in the geographic scale, the sample size and the use of a close vs. distant outgroup, we show the results separately for each dataset as well as pooled results. For all measures of codon preference, there is a weak trend in both species for a higher ratio of unpreferred to preferred (and AT→GC) fixations compared with polymorphism, as expected under a model of translational selection. However, in most cases this trend is not significant, providing little evidence for the action of selection on codon usage in these data. Differences between the ratios for polymorphism and fixation are only significant in the species-wide polymorphism data (Chapter 4) of *A. lyrata*, comparing the ratio of AT→GC vs GC→AT changes (Fisher's exact test $p<0.05$), although it is also close to significance in this same dataset when comparing $P_{mult}\rightarrow U_{mult}$ vs. $U_{mult}\rightarrow P_{mult}$ changes ($p=0.06$).

With the exception of this species-wide *A. lyrata* dataset, both species show a consistent trend towards an excess fixation of preferred to unpreferred and GC→AT changes for both polymorphism and divergence. This excess is significant in the pooled data for most comparisons, suggesting a departure from equilibrium. However, this is not significant for $P_{tRNA}\rightarrow U_{tRNA}$ vs. $U_{tRNA}\rightarrow P_{tRNA}$ divergence for either species, or for *A. lyrata* $P_{tRNA}\rightarrow U_{tRNA}$ vs. $U_{tRNA}\rightarrow P_{tRNA}$ polymorphism (Table 3.3). Thus, while our analysis fits with previous evidence for a departure from equilibrium, it is possible that this results from incorrect assumptions about codon preferences. The use of *Brassica* sequence for parsimony analysis could also lead to biases in the estimation of fixation rate, due to a significant proportion of multiple substitutions. However, analysis of the data using only close outgroup sequence (*A.*

glabra and *Capsella* species) generally shows a similar trend (Table 3.4) of excess $P_{mult} \rightarrow U_{mult}$ and GC→AT, although not $P_{tRNA} \rightarrow U_{tRNA}$.

Table 3.4 Polymorphism and fixation of codon usage in *A. thaliana* and *A. lyrata*.

species	dataset		$U_{mult} \rightarrow P_{mult}$/ $P_{mult} \rightarrow U_{mult}$[1]	$U_{tRNA} \rightarrow P_{tRNA}$/ $P_{tRNA} \rightarrow U_{tRNA}$[2]	AT→GC/ GC→AT[3]
A. thaliana	published	polymorphic	0.7 (43/31)	0.68 (32/47)	0.86 (38/44)
		fixed	0.62 (31/50)*	0.91 (30/33)	0.48 (20/42)*
	Chapter 3	polymorphic	0.08 (1/12)*	0.63 (12/19)	0.29 (5/17)*
		fixed	0.44 (4/9)	0.57 (4/7)	0.52 (11/21)
	close outgroup	polymorphic	0.46 (5/11)	0.42 (5/12)	0.5 (8/16)
		fixed	0.65 (13/20)	1.2 (18/15)	0.68 (15/22)
	pooled	polymorphic	0.58 (32/55)*	0.67 (44/66)*	0.70 (43/61)*
		fixed	0.59 (35/59) *	0.85 (34/40)	0.49 (31/63)*
A. lyrata	Chapter 6	polymorphic	0.39 (7/18)*	0.28 (5/18)*	0.61 (14/23)
		fixed	0.47 (20/43)*	0.61 (17/28)	0.53 (27/51)*
	Chapter 3	polymorphic	0.63 (10/16)	1.43 (10/7)	0.31 (8/26)*
		fixed	2.25 (9/4)	1.0 (6/6)	1 (12/12)
	close outgroup	polymorphic	1.2 (7/6)	0.83 (5/6)	0.64 (7/11)
		fixed	0.48 (10/21)*	0.73 (11/15)	0.57 (17/30)
	pooled	polymorphic	0.5 (17/34)*	0.6 (15/25)	0.45 (22/49)*
		fixed	0.61 (29/47)*	0.68 (23/34)	0.62 (39/45)*

Values given are ratios of change for polymorphism and divergence for unpreferred to preferred relative to preferred to unpreferred changes using 1, multivariate definitions of codon bias and 2, tRNA based definitions and 3, comparisons of the ratio of AT→GC relative to GC→AT changes. Values in parentheses are the observed numbers of each class of change.

*, $p<0.05$ for 1-way χ^2 comparison of U→P (AT→GC) vs. P→U (GC→AT)

It is notable that the species-wide (Chapter 4) *A. lyrata* dataset shows evidence for a more balanced ratio of GC→AT vs. AT→GC fixations relative to polymorphism, suggesting the possibility of an AT→GC fixation bias species-wide. This may result from either the action of natural selection or biased gene conversion favouring GC substitutions. Further comparisons of within- vs. between-population samples should be made to confirm this result. Similarly, it will be important to get a larger number of sequences from closer outgroups, to get a clearer picture of relative

fixation rates. Nevertheless, the overall pattern provides little evidence for either elevated mutation or fixation of AT→GC changes in *A. lyrata*; there are no significant differences between the species in the ratio of changes for either polymorphism or fixation (all Fisher's exact tests $p>0.05$) in the pooled dataset. This leaves the significant elevation of GC_3 in *A. lyrata* unexplained. One possibility is that the ancestor of both species had a higher equilibrium GC content, and that a lower rate of synonymous substitution in *A. lyrata* generates a weak but consistently higher GC_3 compared with *A. thaliana*. Comparisons of GC content at synonymous sites and introns between *A. thaliana* and *Brassica rapa* are in accordance with this hypothesis; *Brassica* shows a strong elevation of GC compared with *A. thaliana* (TIFFIN and HAHN 2002). Furthermore, parsimony-based estimates of the total number of synonymous substitutions using our three species comparisons of single sequences per species show a slightly but significantly elevated number of synonymous changes in *A. thaliana* (583 in *A. thaliana* vs. 514 in *A. lyrata*; $\chi^2=4.34$, $p<0.05$). Finally, the general excess of GC→AT fixations in both species would be consistent with the hypothesis of an overall decline in GC content in both species. It will also be important to extend the analysis of codon bias evolution across more related species, to further test for the relative roles of mutation and selection in the decline in GC content in *Arabidopsis*.

3.3.5 Comparison of replacement divergence and polymorphism

Table 3.4 shows the ratio of amino acid to synonymous (A/S ratio) polymorphism and divergence in the loci surveyed in *A. thaliana* vs. *A. lyrata*. Consistent with previous results (BUSTAMANTE *et al.* 2002; WEINREICH and RAND

2000), the pooled *A. thaliana* published data show a highly significant excess of amino acid polymorphism relative to synonymous polymorphism compared with fixed differences (Fisher's exact test $p<0.001$), and the neutrality index is considerably greater than one (2.6), as expected if a large number of slightly deleterious amino acid mutations are segregating. In contrast, there is no significant difference in *A. lyrata* between the A/S ratio for polymorphism compared with fixed differences, and the neutrality index is in fact slightly less than 1 (0.856). Furthermore, the A/S ratio for polymorphism is significantly different between species (Fisher's exact test $p<0.01$).

However, there is a contrast between the overall published *A. thaliana* data and *A. thaliana* polymorphism data collected by us. Although we have a much smaller dataset, we find an A/S ratio for polymorphism that is much more in line with *A. lyrata* data, and is significantly lower than the ratio from the published dataset (Fisher's exact test $p<0.001$). In contrast, there is no significant difference in the ratio of unpreferred to preferred polymorphisms between the datasets (Table 3.3). The difference in the A/S ratio might simply result from our small sample sizes; most of the published amino acid polymorphisms are segregating at low frequencies (PURUGGANAN and SUDDITH 1998), and it is possible that our small samples failed to pick up these variants. However, the analysis of polymorphism at fragments of our five nuclear loci in much larger samples (Zheng and Nordborg) also shows a significantly reduced A/S ratio for polymorphism compared with published data (Table 3.4; $p<0.05$). This result calls into question the generality of the observed excess amino acid polymorphism in *A. thaliana*, and suggests the

importance of extending this comparison to a more extensive genome-wide sample of loci in both species.

Table 3.5- Polymorphism and divergence of amino acid (A) and synonymous (S) changes in *A. thaliana* and *A. lyrata*.

	dataset	A/S polymorphism	A/S divergence
A. thaliana	published***	0.77 (84/109)	0.30 (32/107)
	Chapter 3	0.07 (2/28)	0.11 (3/27)
	Zheng and Nordborg	0.13 (2/15)	0.13 (1/8)
A. lyrata	Chapter 3	0.29 (15/52)	0.15 (4/26)
	Chapter 6	0.47 (20/43)	0.52 (42/81)
	pooled	0.37 (35/95)	0.43 (46/107)

***, $p<0.001$

3.3.6 Forces driving rates of protein evolution

As in our initial study, we find no evidence for an elevated rate of amino acid substitution in *A. thaliana*; the A/S ratio for fixed differences is very similar in the two taxa (Table 3.4). Furthermore, parsimony analysis using single sequences for each species also indicates that the A/S ratio has remained constant (A/S ratios are 230/583 for *A. thaliana*, 213/540 for *A. lyrata*). Also consistent with previous results, we find a significant negative correlation between EST- estimated gene expression and the rate of protein evolution (K_a vs maximum EST expression: Spearman's r= -0.242, $p<0.05$), but not for the rate of synonymous substitution (K_s vs maximum EST expression: $p>>0.05$). Furthermore, we find a stronger correlation between

MPSS estimates of gene expression level and K_a (K_a vs. maximum MPSS expression: Spearman's r=-0.32, p<0.01), as expected, since these latter data represent more quantitative estimates of gene expression.

If, as hypothesized by Duret and Mouchiroud (DURET and MOUCHIROUD 2000) for human genes, the observed relationship between gene expression and amino acid substitution rate is driven primarily by an effect of higher selective constraints on proteins expressed in a greater diversity of tissues, we would expect that tissue number should show the highest correlation with K_a, and that there would be no residual effect of expression level when correcting for tissue number. The MPSS data on gene expression from five tissues allows for a detailed test of this prediction. Because of the potential for errors in the MPSS dataset for values less than 10 parts per million (B. Meyers, pers comm), we only consider a gene expressed in a given tissue if it has values greater than or equal to 10. Using this measure, we find a stronger negative correlation between K_a and tissue number (Spearman's r=-0.384, p<0.001) than with estimates of gene expression level. Furthermore, a partial correlation shows no significant effect of either EST or MPSS expression level once tissue number is controlled for ($p>>0.05$), suggesting that the diversity of tissues in which a gene is expressed may be the primary factor driving the observed effect on protein evolution, and that the specificity of gene expression may be a general factor driving variation in rates of protein evolution in *Arabidopsis*.

Chapter 4

Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*

4.1 Introduction

High levels of inbreeding are expected to be associated with a strong reduction in the effective size and recombination rate of natural populations. These effects reduce levels of diversity within natural populations, increase linkage disequilibrium, and reduce the effectiveness of natural selection on weakly selected mutations. Most models of the effects of inbreeding on patterns of polymorphism and divergence have examined single populations at equilibrium that differ only in breeding system. However, species may differ in other ways in addition to the breeding system, and the assumptions of the single-population equilibrium models may be violated in various ways. For example, in the presence of population subdivision, the interaction between inbreeding and natural selection will have a relatively minor effect on the amount of diversity in species-wide samples (CHARLESWORTH *et al.* 1997), and patterns of genetic diversity and linkage disequilibrium will also be influenced by historical rates of gene flow and population extinction (WAKELEY and ALIACAR 2001). Furthermore, recent changes in population size or migration patterns can lead to departures from neutral equilibrium, which will also influence both patterns of diversity (TAJIMA 1989a) and linkage disequilibrium (*e.g.* PRITCHARD and PRZEWORSKI 2001). Another important difference in inbreeding populations may be the evolution of higher rates of

crossing-over. This has been observed in a number of comparisons between related selfing and outcrossing plants (reviewed in CHARLESWORTH and CHARLESWORTH 1979), and could be a general evolutionary trend that would partly counteract the reduction in effective recombination rates in inbreeders.

While the polymorphism patterns in *A. thaliana* show some striking differences from those found in the outcrossing *D. melanogaster*, it is difficult to know which of the many differences between these species may be important. Comparisons with closely related self-incompatible species should be much more helpful in understanding what influences levels and patterns of diversity, and should help illuminate the joint effects of mating system, natural selection and life history. For example, plant populations may be expected often to show strong population structure, and it is unclear to what degree patterns observed in *A. thaliana* may simply reflect effects of population subdivision in a widely dispersed species (SAVOLAINEN *et al.* 2000). In addition, climate changes associated with recent glaciation events may have had a general influence on the amount and structuring of genetic diversity in related plant populations (COMES and KADEREIT 1998). Distinguishing the influences of breeding system from other factors requires a detailed analysis of polymorphism patterns within and between populations from both highly inbreeding species and their close relatives.

In this study, we analyse DNA sequence polymorphism at six loci to assess the patterns of genetic diversity within and between four populations of *Arabidopsis lyrata*. The observed patterns are compared with species-wide polymorphism at orthologous loci, as well as within- and between-population patterns at other studied loci in *A. thaliana*. *A. lyrata* has a disjunct circumpolar distribution in Northern

Europe, North America and Japan, and is self-incompatible and thus highly outcrossing in natural populations. A survey of DNA sequence polymorphism (SAVOLAINEN *et al.* 2000) provided preliminary evidence for higher levels of within-population diversity at the ADH locus than those estimated within *A. thaliana* populations based on RFLP data (BERGELSON *et al.* 1998), though species-wide polymorphism was lower in *A. lyrata*, and there was strong population subdivision in the outcrossing species. There was also evidence for a significant departure from neutral equilibrium expectations at the ADH locus in North American *A. lyrata* populations, with an excess of intermediate frequency variants, suggesting the possibility of a recent bottleneck. Further investigation of polymorphism patterns at multiple loci are necessary to understand the relative importance of natural selection and demographic history in structuring genetic diversity in this outcrossing species.

4.2 Methods

Loci Surveyed

Table 4.1 shows the genes used in our study, their locus names from the *A. thaliana* genome project, and their map positions, obtained from The Arabidopsis Information Resource (TAIR-www.arabidopsis.org). The sampled genes represent five unlinked nuclear loci and a single locus (RBCL) from the chloroplast. Two loci (PGIC and CAUL) have been previously sequenced in multiple *A. thaliana* ecotypes

Table 4.1- Genes surveyed in this study and their genomic positions in *A. thaliana*

Locus	Locus name	Map position (Chromosome:cM)	Size of aligned region (b.p.)
CAUL	AT1G26310	1:35	862
ETR1	AT1G66340	1:92	2111
HAT4	AT4G16780	4:36	726
PgiC	AT5G42740	5:85	1477
ScADH	AT4G05530	4:23	1249
RBCL	-	Chlor: 54958-56397	1297

(KAWABE *et al.* 2000; PURUGGANAN and SUDDITH 1998). For these loci, we present the results based on our smaller sample of ecotypes, as well as the results for the total sample for the gene regions that were studied in our sequencing survey.

Population samples

A total of 17 *A. lyrata* individuals from independent field-collected maternal families were used for this study. The samples include two populations from the European subspecies *petraea* (four individuals from Scotland and five from Iceland, collected by R. Ennos and M. Schierup, respectively) and two from the North American subspecies *lyrata* (four individuals from Michigan and four from North Carolina, collected by C. Langley and R. Mauricio). Seeds from a worldwide collection of *A. thaliana* ecotypes were also obtained from the Nottingham Arabidopsis Stock Centre and the Arabidopsis Biological Resource centre. For this study, we used single individuals from the ecotypes COL, LER, NAA, BAY, B1, AN-O, KAS, PUZ, KNO, CIBC and GOT, as well as an individual from a local

Edinburgh population of *A. thaliana* (GB). For several loci the *A. thaliana* genome project-derived sequence from the COL ecotype was used directly in the analyses; these sequences did not contribute any singleton variants, supporting the accuracy of these sequence data.

DNA Extraction, PCR, and Sequencing

Genomic DNA was isolated from leaves using the method of Ingram et al (INGRAM *et al.* 1997). PCR primers were designed using the *A. thaliana* sequence information, and amplified in both species for 30 cycles consisting of 1 min. denaturing at 95^0C, 30 seconds annealing at 55^0C, and 2 min. extension at 72^0C. PCR primer sequences are: Sc-ADH-F 5'GGCATTCCTCCAGCGAC3', Sc-ADH-R 5'CTTCCGTCGTCGTCTCTTC3'; HAT4-F 5'CGTGAACAGACCACCGTC3', HAT4-R 5'AGCGTCAAAAGTCAAGCCGT3'; ETR1-F 5'-AGACCAAAGCTCATGCATTTCT-3', ETR1-R 5'-TGTTGACTCATGAGATTAGAAGCA-3'; RBCL-F 5'-TGTCACCACAAACAGAGACTAA, RBCL-R 5'-CTTTCCATACTTCACAAGCAG-3'; PGIC-F 5'-TGCTGTSAGCACTAATCTTGCG-3'; PGIC-R 5'-TCGAACCCGGGAGAGGTAGACCA-3'; CAUL-F 5'-GTACTCGAACGCTACGAGAGGT-3', CAUL-R 5'-GCATGCTGTTTTCCTCCTGT -3'. For *A. thaliana* sequences and the chloroplast locus RBCL, we sequenced PCR products directly on both strands. For *A. lyrata* nuclear genes, PCR fragments were cloned into the PCR 2.1 vector using the TA cloning kit (Invitrogen Life Technologies), and at least four clones were sequenced

individually using the ABI automated sequencing facilities at the University of Edinburgh Institute of Cell, Animal and Population Biology. In as many cases as possible, we sequenced both alleles from heterozygous individuals, and all variant alleles included in final analyses were confirmed using multiple clones. Wherever possible, we also confirmed putative homozygous individuals using direct sequencing, and included both copies in the sample analysed. For loci and populations with few indel polymorphisms, direct sequencing was also used to confirm polymorphic sites in heterozygotes. PCR primers, the Universal M13 primers, as well as several internal sequencing primers were used in sequencing. In all cases, direct sequencing from homozygous individuals and/or sequence analysis of a large number of clones (approximately 10 per individual) from a small subset of individuals indicated that the genes were single copy in *A. lyrata*, with one or two haplotypes identified per individual, with the rare exception of a small number of obvious PCR recombinants. All new sequences are deposited into Genbank, with accession numbers AY174502-AY174656.

A. thaliana polymorphism datasets

Data on nucleotide polymorphism within four *A. thaliana* populations (n=63 to 91) at six loci (98 to 413 silent sites) were obtained from E. Stahl (E. A. Stahl, R.A. Hufft, M. Kreitman, J. Bergelson, submitted manuscript). These loci include four random regions from effectively unlinked genomic clones along chromosome 3, as well as regions from the genes Ap3 and CAUL. Note that the sampled *A. thaliana* populations were chosen because they were known to be polymorphic at the Rpm1 locus and had persisted over several field seasons. *A. thaliana* polymorphism

datasets using worldwide samples were also extracted from Genbank at http://www.ncbi.nlm.nih.gov/Entrez/. The loci examined from published datasets were CAL (PURUGGANAN and SUDDITH 1998), fah1 and f3h (AGUADE 2001), AP3 and PI (PURUGGANAN and SUDDITH 1999), LFY TF1 and AP1 (OLSEN *et al.* 2002), ADH (INNAN *et al.* 1996), CHIA (KAWABE *et al.* 1997), CHIB (KAWABE and MIYASHITA 1999), PGIC (KAWABE *et al.* 2000), and CHI (KUITTINEN and AGUADE 2000). We also included ACL5, submitted directly to Genbank by K. Yoshida, T. Kamiya, A. Kawabe and N. T. Miyashita and MYB, submitted by T. Kamiya, A. Kawabe and N. T. Miyashita.

Data Analysis

Sequences were aligned using Clustal W (THOMPSON *et al.* 1994), and modified by eye using Genedoc (NICHOLAS 1997). For each locus, two estimators of the population mutation parameter per nucleotide site $\theta=4N_e u$ (where *Ne* is the effective population size and u is the mutation rate) were calculated, the average number of pairwise differences π (TAJIMA 1989b), and Watterson's estimator θ_w (WATTERSON 1975), based on the number of segregating sites. The average within-population diversity was also calculated for each subspecies, weighted by the sample sizes as:

$$\pi_{within} = \frac{\pi_1 n_1 + \pi_2 n_2}{n_1 + n_2}$$

where π_1 and n_1 are the average pairwise differences and sample size for population 1, and π_2 and n_2 are the average pairwise differences and sample size for population 2.

Average pairwise divergence at silent sites (D_{sil}) between the species was also estimated, excluding unalignable noncoding regions from several loci. We considered silent sites (introns and synonymous) and replacement sites separately for polymorphism analysis, and excluded insertion-deletion polymorphisms (indels). Tajima's D (TAJIMA 1989b) statistic was used to test the neutral prediction that π and θ should be equivalent. This test measures the skew in the frequency spectrum; a negative D value indicates an excess of rare polymorphisms, while a positive D suggests excess intermediate frequency variants. Significance of Tajima's test was calculated assuming no recombination, which is conservative for testing departures from neutral equilibrium. Multilocus significance of the mean and variance of Tajima's D was assessed using the HKA program of J. Hey (http://lifesci.rutgers.edu/~heylab/DistributedProgramsandData.htm#HKA). Tests of population differentiation between all pairwise comparisons of the four populations, all populations together, and between the two subspecies, were applied using two statistics, K_{st}* (HUDSON *et al.* 1992a) and S_{nn} (HUDSON 2000), chosen because they have been shown to have the highest power to detect population differentiation under simple models of population structure (HUDSON 2000), and are likely to measure different aspects of differentiation. S_{nn} calculates the proportion of nearest neighbours in sequence space that are found in the same population, while K_{st}* is a weighted measure of the ratio of the average pairwise differences within populations to the total average pairwise differences. Significance levels for these statistics were assessed using permutation tests, with 1000 permutations. We also estimated the population differentiation statistic F_{st}, using the method of Hudson et al. (HUDSON *et al.* 1992b). Note that under a symmetric island model of migration, $F_{st} = 1/(1+4Nm)$,

where m is the migration rate, but under more complicated migration models, the relationship between F_{st} and migration rates is often unclear (e.g. WHITLOCK and MCCAULEY 1999). All the above analyses were performed using DnaSP v. 3.95 (ROZAS and ROZAS 1999). A multilocus HKA test (HUDSON *et al.* 1987) was used to test the significance of the ratio of polymorphism within *A. lyrata* to divergence between *A. lyrata* and *A. thaliana* for silent sites, using all six loci. To test for significance, 10,000 coalescent simulations were run, using the HKA program.

In order to get a multilocus estimate of θ_w, and to test the significance of differences between species and subspecies, we used a simple maximum likelihood-based estimate using coalescent theory. The likelihood of θ given S observed segregating sites in a sample of size n can be calculated using the recursion equation given by Hudson (1990, equation 12), which gives the probability of observing S segregating sites in a sample of size n, for a given value of θ:

$$L_n(\theta / S) \propto P_n(S/\theta) = \sum_{i=0}^{S} P_{n-1}(S-i/\theta) Q_n(i/\theta)$$

where

$$P_2(S/\theta) = \left(\frac{\theta}{1+\theta}\right)^S \frac{1}{1+\theta}$$

and

$$Q_n(S/\theta) = \left(\frac{\theta}{\theta+n-1}\right)^S \frac{n-1}{\theta+n-1}$$

The likelihoods of a range of θ values for each locus were calculated using this equation, given the observed numbers of segregating silent sites (S), and the sample size (n). To correct for differences in sequence lengths among loci, we calculated the likelihood for each locus for a range of possible values of θ per base pair, from 0 to

0.05. The multilocus likelihood of θ per base pair was then calculated as the product of the likelihoods for each locus. Significant differences among taxa for θ were then assessed using the χ^2 approximation. This approach assumes a fixed mutation rate across loci, no recombination within loci, and free recombination between loci. Given the distance between all loci based on the *A. thaliana* genome project (*i.e.* at least several megabases), the assumption of free between-locus recombination is reasonable, even for the highly selfing *A. thaliana* (see NORDBORG *et al.* 2002). Because of the assumption of a single underlying mutation rate, we exclude the RBCL locus for this analysis, since chloroplast loci are known to have much lower mutation rates than nuclear genes (WOLFE *et al.* 1987, see results). For our analyses, the χ^2 credibility interval serves as an approximate assessment of the likelihood surface, and marginally significant differences should be treated with caution. Nevertheless, several considerations suggest that our tests for significance should be conservative. First, the assumption of no within-locus recombination should not affect the maximum likelihood estimate of θ (θ_{MLE}), but will inflate the credibility intervals, and is therefore conservative for testing for differences among taxa. This expectation has been confirmed using simulated data assuming mutation and recombination parameters relevant to this dataset (see Appendix A.2). Furthermore, since the two *A. lyrata* subspecies will have shared coalescent histories, differences in N_e between populations are likely to have less effect on $\hat{\theta}$ than if the populations were independent, forcing diversity levels to be more similar than assumed under the χ^2. Further implications of violations the assumptions of the standard neutral model will be discussed below.

Several summary statistics of the amount of linkage disequilibrium were used. The number of haplotypes (h) and the minimum number of recombination events (R_m) (HUDSON and KAPLAN 1985) and the B statistic (WALL 1999) were calculated using DnaSP v. 3.95. In addition, we estimated the population recombination parameter $\rho=4N_e r$, where r is the per base pair recombination rate, using the method of Wall (WALL 2000), which uses coalescent simulations to estimate the maximum likelihood of ρ given h and R_m. Maximum-likelihood-based estimates of ρ from polymorphism data will be referred to in the results as $\hat{\rho}$. Coalescent simulations were run using the standard neutral model, which assumes a single constant-sized population at equilibrium. While estimates of ρ should thus be treated with caution, comparison with an independent expectation of ρ provides a test of the fit of the standard neutral model to the data (see below). Note that for a partially self-fertilizing species, the coalescent theory assuming the standard neutral model predicts that the effective recombination parameter $\hat{\rho}_s = \rho$ (1-s), where s is the selfing rate (NORDBORG 2000a). Approximate 95% credibility intervals for $\hat{\rho}$ were obtained using the 2-unit χ^2 approximation, with 10,000 to 100,000 coalescent simulations run for each locus-population combination. Coalescent simulations were run to test the significance of Wall's B and $\hat{\rho}$, as described below.

If there is independent information on rates of recombination at individual loci, the amount of linkage disequilibrium from population samples can be compared to the amount expected, as a test of the standard neutral model (e.g. ANDOLFATTO and PRZEWORSKI 2000; FRISSE *et al.* 2001). We compared

polymorphism-based estimates of $\hat{\rho}$ and B to values expected based on levels of polymorphism and physical estimates of recombination rate at each locus using *A. thaliana* mapping data, and assuming that recombination rates have remained constant since species divergence. Under the standard neutral model, N_e can be estimated from θ and an estimate of the mutation rate, using the equation given above. Using map-based estimates of the physical rate of recombination (r_{map}) for each locus (see below), the expected ρ for an outcrossing population can then be calculated by:

$$E(\rho) = \frac{\hat{\theta}}{u} r_{map}$$

For a partially self-fertilizing species at equilibrium with selfing rate s, the expectation of the estimate of ρ becomes:

$$E(\rho_s) = \frac{\hat{\theta}}{u} r_{map}(1-F)$$

where F is the inbreeding coefficient (HAGENBLAD and NORDBORG 2002; NORDBORG 2000a); at inbreeding equilibrium, $F=s/(2-s)$. For *A. thaliana*, we assume a lower-bound estimate of the selfing rate of 96%, based on a DNA sequence diversity study from large North American populations (E.A. Stahl R. Hufft, M. Kreitman, J. Bergelson, manuscript submitted). For each population studied, E(ρ) was calculated using multilocus maximum likelihood estimates of θ from the number of segregating sites, combined with locus-specific estimates of r_{map}. The per-generation mutation rate for nuclear genes (u) is unknown for the study species, but upper- and lower- bound estimates of u can be used for the E(ρ) calculations. In particular, we used two estimates of the mutation rate; we first assumed the

substitution rate of (KOCH *et al.* 2000), who used fossil pollen data from the Brassicaceae and estimated the mutation rate as 1.5×10^{-8} per site per generation. This is more than twice the estimate of the plant substitution rate of 6.5×10^{-9} based on the proposed monocot-dicot divergence time (WOLFE *et al.* 1987), which we take as our lower-bound; this value is comparable to average estimates of substitution rates in grasses using the ADH locus ($6.0\text{-}7.0 \times 10^{-9}$, (GAUT *et al.* 1996).

We estimated the rate of recombination (r_{map}) at each locus by using information from the *A. thaliana* genome project. First, map-based estimates of recombination rates across all five chromosomes were determined using the marker position information from TAIR. Briefly, physical positions were recorded for genetic markers that have both been mapped to the Recombinant Inbred (RI) recombination map (see http://nasc.nott.ac.uk/new_ri_map.html), and have been precisely physically mapped based on flanking sequences (www.arabidopsis.org). To estimate rates of crossing over at each locus, the relationship between genetic and physical distance was examined for each chromosome. Because marker density is low close to the centromere, each chromosome was divided into the two arms, with the boundary representing the centromere, which lacks markers. Each arm had an average of 40 markers, with a total of 400 markers used genome-wide. Third-order polynomials were fitted to the relationship between genetic and physical distance (in cM/Mb) in each arm, which is similar to the approach of (KLIMAN and HEY 1993). Rates of crossing over were then estimated by taking the first derivative of this polynomial at the physical position of each locus of interest. With two exceptions, the polynomial fits explained greater than 97% of the variance in genetic

distance; the right arm of chromosome 5 had an R^2 of 0.96, while the left arm of chromosome 2, represented by 19 markers, had an R^2 of 0.87.

For significance testing of the B statistic, 10,000 coalescent simulations were run for each locus, using E(ρ) as the recombination parameter, and conditioning on the observed number of segregating sites (S) and sample size (n). Significance of the observed B values was assessed as the 1-tailed probability of observing a value as high or higher than $B_{observed}$. For h and Rm, 100,000 simulations were run given E(ρ), using the program of Wall (2000), conditioning on the observed number of segregating sites S and sample size n. Approximate assessment of significant departure was calculated as the sum of the probabilities as low or lower than that for the observed h and Rm.

4.3 Results

4.3.1 Patterns of DNA polymorphism:

Figure 4.1 shows the polymorphic sites for all loci from *A. lyrata*, and Table 4.2 summarizes the patterns of nucleotide polymorphism within the three taxa. Species-wide levels of silent polymorphism for the nuclear loci in *A. lyrata* are high; total nucleotide site diversities (π) for silent sites for the entire sample range from 0.01 (CAUL) to 0.05 (PGIC). The chloroplast locus RBCL has only one segregating site in *A. lyrata*, and an estimate of π_{sil} approximately an order of magnitude smaller than the nuclear genes (0.002). This is consistent with a large reduction in the mutation rate in the chloroplast genome, and there is also correspondingly low silent divergence between *A. thaliana* and *A. lyrata* at this locus (Table 4.2).

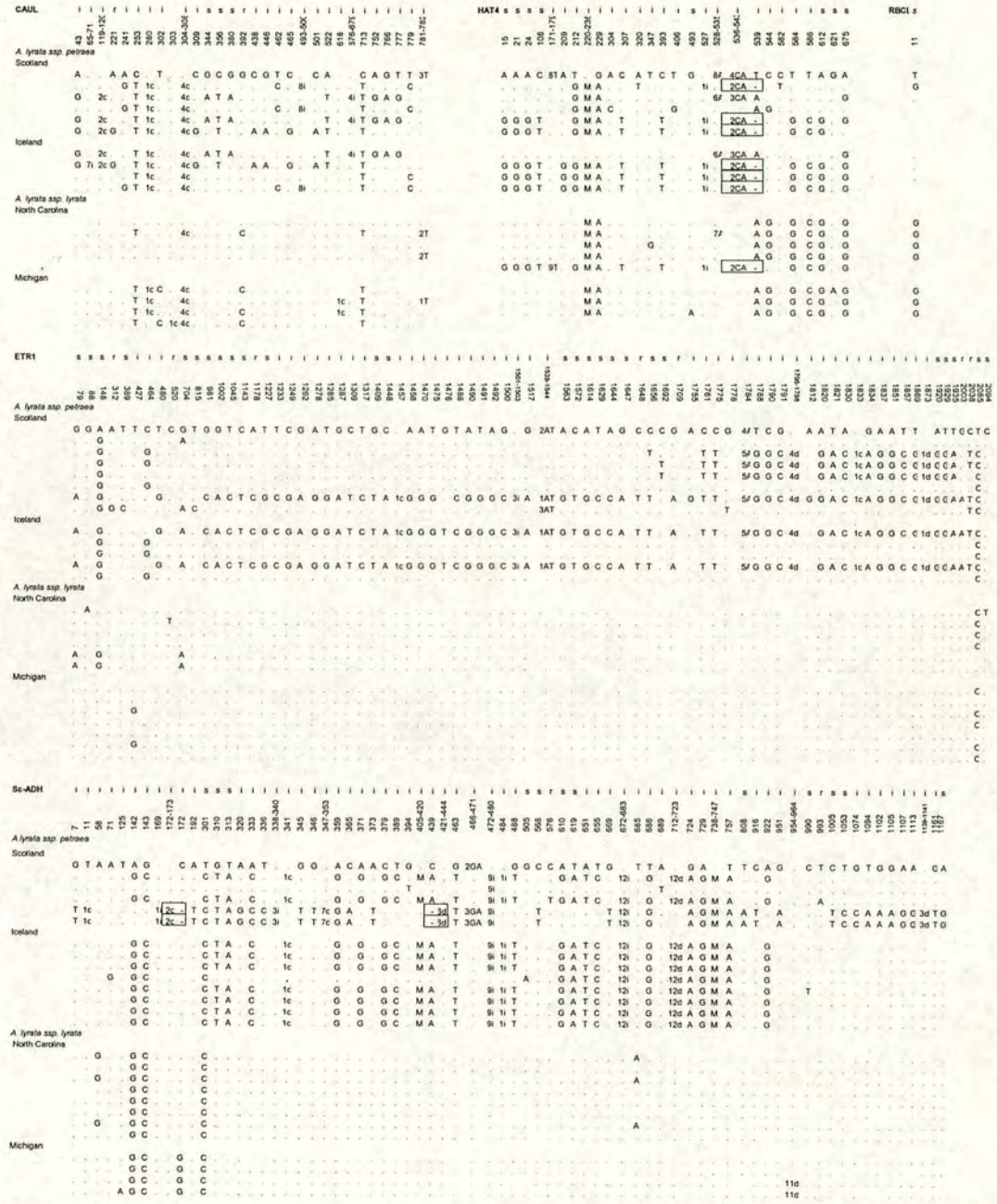


Figure 4.1 Polymorphic sites for the 6 loci surveyed in *Arabidopsis lyrata*. Top line indicates the location of the substitution (i=intron, s=synonymous, r=replacement). Positions of polymorphic sites within the alignment are given in the next line. Indels are shown as i (insertion relative to the first sequence) and d (deletion relative to the first sequence), with the number of base pairs given. Boxes represent segregating sites in indels. Microsatellite polymorphisms are given as the number of repeats and the type of nucleotide.

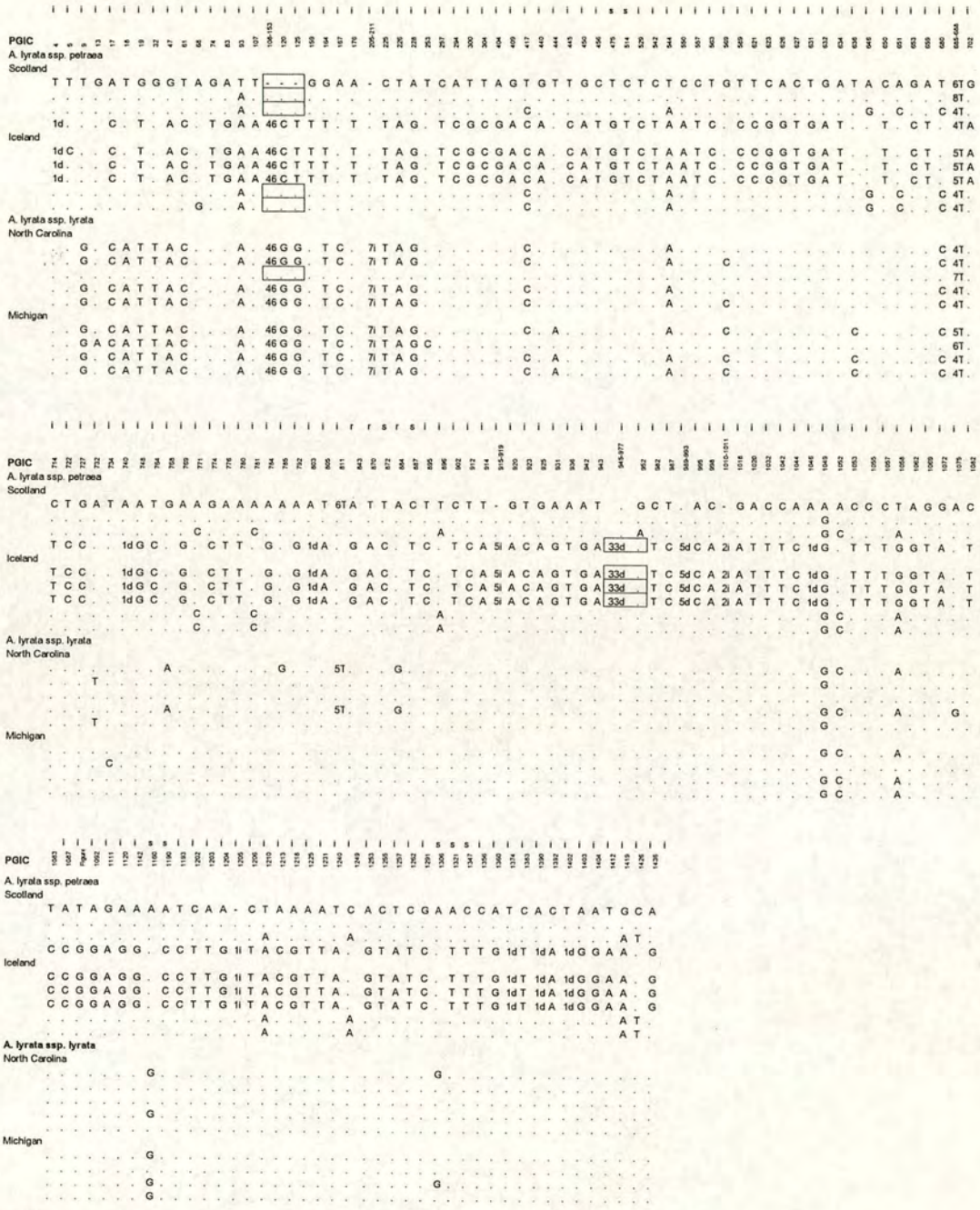


Figure 4.1 continued

Amino acid polymorphism is generally much lower than silent site variability, with π_{rep}/π_{sil} ranging from 0 for Hat4 to 0.3 for CAUL, consistent with the action of strong purifying selection on the majority of nonsynonymous sites. All loci also have polymorphic insertions and deletions (indels- Figure 4.1), all of which are within introns. The ratio of indels to nucleotide substitutions in *A. lyrata* introns ranged from 0.09 in PGIC to 0.56 in CAUL, with an average of 0.3.

Under the standard neutral model, levels of polymorphism within species and numbers of fixed differences between species should be correlated across loci, since both depend on the mutation rate. Departures from this prediction can be caused by the action of natural selection at individual loci, which can obscure any general assessment of the neutral patterns of variability using multilocus data. The HKA test can be performed to test this neutral prediction (HUDSON *et al.* 1987). For *A. lyrata ssp. petraea* and *A. lyrata ssp. lyrata*, we observe no evidence for a significant deviation among loci in the ratio of polymorphism to fixed differences (*ssp. petraea* d.f.=5,χ^2=4.82, p>>0.05; *ssp. lyrata* d.f.=5, χ^2=10.21, p>0.05), consistent with neutral expectations. In contrast, *A. thaliana* shows a significant multilocus HKA (d.f.=5, χ^2=14.38, p<0.05). For *A. thaliana*, the deviation from neutrality appears to be mostly contributed by the HAT4 locus, which shows unusually high levels of polymorphism; θ for this locus is higher than any other gene sampled from *A. thaliana*, with the exception of RPM1, which is thought to be under long-term balancing selection (STAHL *et al.* 1999). If we exclude HAT4, the HKA test for *A. thaliana* is no longer significant (d.f.=4, χ^2=4.7, p>>0.05).

Table 4.2- Summary of silent nucleotide polymorphism.

Locus	D_{sil}[a]	*A. lyrata ssp. petraea*					*A. lyrata ssp. lyrata*					*A. thaliana*[†]			
		L[b]	n	S[c]	π_{within}[d]	π_{total}[e]	L[b]	n	S[c]	π_{within}[d]	π_{total}[e]	L[b]	n	S[c]	π_{total}[e]
CAUL	0.131	567	10	18	0.0150	0.0135	565	8	3	0.0018	0.0027	587	8 (22)	8 (22)	0.0042 (0.0049)
ETR1	0.144	874	13	60	0.0284	0.0276	874	13	7	0.0021	0.0022	813	9	30	0.0192
HAT4	0.107	421	10	18	0.0166	0.0183	428	8	11	0.0064	0.0064	417	5	28	0.0302
PGIC	0.136	1100	10	135	0.0653	0.0625	1144	9	31	0.0082	0.0087	1115	4 (24)	2 (34)	0.0010 (0.0037)
SCADH	0.117	703	14	46	0.0175	0.0206	761	12	4	0.0012	0.0019	920	5	18	0.0087
RBCL	0.017	319	7	1	0.0009	0.0013	321	7	0	0.0000	0.0000	318	9	1	0.0012

[a], average pairwise silent divergence
[b], number of silent sites (bp)
[c], number of segregating silent sites
[d], average within-population nucleotide site diversity, per bp
[e], total nucleotide site diversity, per bp
[f], parentheses show values from total sample, including individuals surveyed in other studies

4.3.2 Among-taxa differences in levels of polymorphism

Except for HAT4, species-wide silent nuclear locus polymorphism levels for *A. lyrata ssp. petraea* are all higher than those for the *A. thaliana* orthologues (Table 4.2). There is also a consistent difference in polymorphism between the two *A. lyrata* subspecies; for all loci, the European subspecies *A. lyrata ssp. petraea* has much higher levels of both average within-population and total diversity than the North American *A. lyrata ssp. lyrata* (Table 4.2, Figure 4.1). Joint maximum likelihood estimates of θ using silent diversity from the five nuclear loci give values of 0.0047 for *A. lyrata ssp. lyrata*, but 0.0226 for *A. lyrata ssp petraea*. This difference is significant, and suggests a greater than four-fold difference in polymorphism between the two subspecies (Figure 4.2). The joint maximum likelihood estimate of θ for *A. thaliana* is 0.0114, intermediate between the two *A. lyrata* subspecies, but not significantly different from either subspecies. If we estimate θ_{sil} including all available polymorphism studies in *A. thaliana* (excluding Rpm1), we get the nearly identical MLE of 0.0112, and a marginally significant difference from both *A. lyrata* subspecies (Figure 4.2). In *A. lyrata,* the PGIC locus shows very high levels of polymorphism in both subspecies (Table 4.2), and might be under long-term balancing selection (B. Lauga and D. Charlesworth, manuscript in preparation), as has also been suggested in *Arabidopsis thaliana* (KAWABE *et al.* 2000) and *Leavenworthia stylosa* (FILATOV and CHARLESWORTH 1999; LIU *et al.* 1999). Excluding PGIC, θ_{sil} is estimated as 0.0035 in subspecies *lyrata*, and 0.019 in subspecies *petraea*. Similarly, if we exclude the HAT4 locus, θ_{sil} in *A. thaliana* from four nuclear loci in our dataset is 0.0071, or 0.0108 using all available loci, both still intermediate between the two *A. lyrata* subspecies.

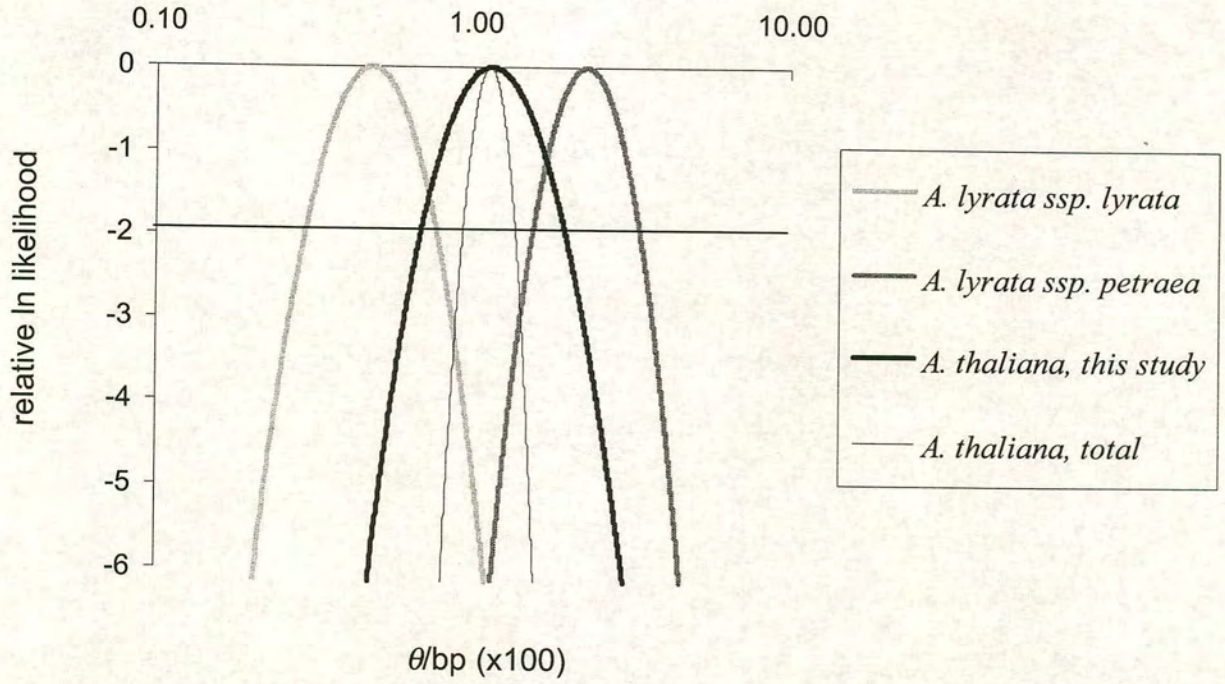


Figure 4.2 Joint maximum likelihood estimates of species-wide θ from the five nuclear loci.

In contrast with species-wide estimates of polymorphism, average levels of within-population diversity are lower in *A. thaliana* than in either *A. lyrata* subspecies (Figure 4.3). Nevertheless, the credibility intervals for θ in Michigan overlap with those for all four *A. thaliana* populations, and within-population diversity is significantly lower in *ssp. lyrata* compared with *ssp. petraea*. The unweighted average of θ_{MLE} within populations is 0.025 for *A. lyrata ssp petraea*, 0.0036 for *A. lyrata ssp lyrata*, and 0.0018 for *A. thaliana*, giving a ratio of within-population diversity among taxa of 14:2:1. Removing PGIC from the *A. lyrata* dataset, the average θ_{MLE} within populations is 0.019 for *A. lyrata ssp petraea* and 0.0031 for *A. lyrata ssp. lyrata*, giving a ratio of 10.5:1.7:1.

4.3.3 Levels of population differentiation

Table 4.3 summarizes tests of population differentiation and estimates of F_{st} between subspecies and subpopulations. Note that marginally significant tests should be treated with caution, given the number of tests performed. Most loci show significant differentiation among all populations, and also between the two subspecies, based on permutation tests. Similarly, most of the pairwise comparisons using single populations from each subspecies show evidence of differentiation, with intermediate values of F_{st}. In contrast, significant differentiation among subpopulations within subspecies is infrequent; only two loci for *A. lyrata ssp. lyrata*, and one locus for *A. lyrata ssp. petraea* are significant using K_{st}^{*}, and only one additional locus (CAUL) is marginally significant for *A. lyrata ssp. petraea* using S_{nn}.

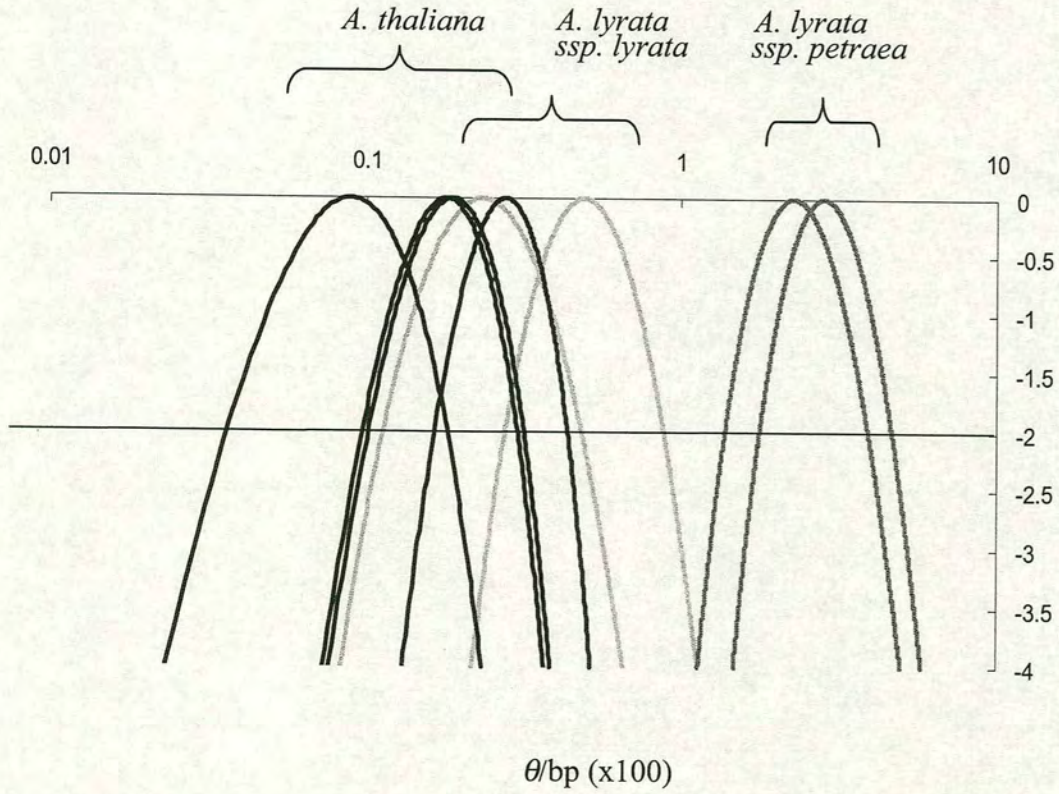


Figure 4.3 Maximum likelihood estimates of within-population θ for *A. lyrata ssp. lyrata*, *A. lyrata ssp. petraea*, and *A. thaliana*.

The single segregating site in RBCL is polymorphic only in Scotland, and shows a fixed difference between Iceland and the North American populations, generating high estimates of differentiation. Only the SCADH locus shows significant differentiation among all pairwise comparisons, and it also shows the highest values of F_{st} for all comparisons except RBCL. Some of the F_{st} values between Scotland and Iceland are negative, likely reflecting a low level of differentiation combined with small within-population samples; small samples can inflate estimates of π_{within} compared with $\pi_{between}$ estimates from larger samples, generating negative values. In contrast, estimates of differentiation between North Carolina and Michigan are larger and positive, although most are not significant. Populations with low levels of polymorphism will have higher estimates of F_{st} and population differentiation (CHARLESWORTH 1998), so that the different F_{st} values do not necessarily imply that within-subspecies migration rates differ between North America and Europe. More than 80% of segregating sites in the sample from subspecies *lyrata* are unique to one or other of the two subpopulations, in contrast to subspecies *petraea*, where less than 30% of polymorphic sites are unique (Table 4.4). There are very few fixed differences, either within or between subspecies, suggesting that, if geographic isolation is complete among any of the subpopulations, it is very recent.

Table 4.3-Tests of geographic differentiation among *A. lyrata* populations

pair		CAUL	ETR1	HAT4	PGIC	SCADH	RBCL
total	F_{st}	**0.163**	**0.181**	**0.335**	**0.280**	**0.536**	**0.727**
	K_{st}*		**	*	**	***	**
	S_{nn}		***	**	**	***	***
ssp pet-ssp lyr	F_{st}	**0.254**	**0.278**	**0.321**	**0.352**	**0.564**	**0.833**
	K_{st}*	**	***	**	***	***	**
	S_{nn}	**	***	**	**	***	***
between ssp.							
MI-IC	F_{st}	**0.205**	**0.285**	**0.577**	**0.503**	**0.903**	**1.000**
	K_{st}*	*	***		*	**	*
	S_{nn}		***	*	***	***	***
MI-SC	F_{st}	**0.267**	**0.280**	**0.426**	**0.215**	**0.343**	**0.500**
	K_{st}*	*	**	*	*	*	
	S_{nn}	*	*	***	***	***	***
NC-IC	F_{st}	**0.238**	**0.241**	**0.342**	**0.461**	**0.886**	**1.000**
	K_{st}*	*	*		**	***	*
	S_{nn}		***	*	*	***	***
NC-SC	F_{st}	**0.242**	**0.232**	**0.249**	**0.100**	**0.345**	**0.500**
	K_{st}*			*		***	
	S_{nn}		*	*		***	***
within ssp.							
MI-NC	F_{st}	**0.492**	**0.124**	**0.000**	**0.077**	**0.607**	-
	K_{st}*				*	**	-
	S_{nn}				*	***	-
SC-IC	F_{st}	**-0.235**	**-0.054**	**0.163**	**-0.007**	**0.287**	**0.000**
	K_{st}*					**	
	S_{nn}		*			**	

*, p<0.05, **, p<0.01, ***, p<0.001

Table 4.4- Assortment of polymorphisms within and between *A. lyrata* populations[a].

populations	shared	fixed	exclusive, pop1	exclusive, pop2
Michigan-North Carolina	13 (0.18)[b]	1 (0.01)	14 (0.19)	45 (0.62)
Scotland-Iceland	253 (0.77)	0 (0)	70 (0.23)	7 (0.02)
ssp lyrata-ssp. petraea	37 (0.10)	0 (0)	33 (0.09)	300 (0.81)

[a], includes indels and microsatellites, where polymorphic microsatellites are counted once, regardless of allele number

[b], proportion of polymorphisms in each category are shown in parentheses

4.3.4 Levels of linkage disequilibrium and tests of the standard neutral model

The patterns of diversity in *A. lyrata* at many of the loci suggest strong haplotype structure; most of the sequences can be grouped into a few sequence classes, with few intermediate types (Figure 4.1). This indication of strong linkage disequilibrium across up to 2000 base pairs of sequence is surprising, given that recombination events should be frequent in an outcrossing species.

Table 4.5 summarizes levels of linkage disequilibrium and neutrality tests for the five nuclear loci in *A. lyrata*. Because of the small number of informative sites for most loci within *A. lyrata ssp. lyrata* subpopulations, we pooled the Michigan and North Carolina populations. The observed number of haplotypes is often small in comparison with the sample size in *A. lyrata*, even for subspecies *petraea*, which has large numbers of segregating sites. Most of these values are non-significant assuming no recombination; haplotype tests assuming no recombination (STROBECK 1987) are significant only for the ETR1 locus in Iceland, where only two haplotypes are observed with 57 segregating sites ($p<0.01$), PGIC in *A. lyrata ssp. petraea*, and the SCADH locus from the total dataset ($p<0.05$), reflecting the strong population structure in the pooled sample for this locus. However, there is strong evidence for

Table 4.5- Levels of linkage disequilibrium and tests of neutrality across the five nuclear loci in *A. lyrata*.

locus	population	n	h	R_m	r_{map}[a]	E(ρ)[b]	$\hat{\rho}$ **(95% C.I.)**[c]	p (h,Rm)[d]	B	p (B)[e]	Tajima's D
CAUL	*A. lyrata ssp. lyrata*	8	4	0	**5.2**	8	2 (0,>500)		0		1.495
	A. lyrata ssp. petraea	10	5	0		39	0 (0,10)	**	0.222		0.896
	SC	6	4	0			0 (0,33)		0.222		0.309
	IC	4	4	0			0 (0,>500)		0.250		-0.117
	A. lyrata	18	8	0		34	0 (0,7)	**	0.200		0.100
ETR1	*A. lyrata ssp. lyrata*	13	6	0	**4.8**	21	0 (0,41)		0		-0.791
	A. lyrata ssp. petraea	13	9	3		98	8 (2,30)	*****	0.677	****	1.086
	SC	8	8	2			8 (2,39)	**	0.719	****	-0.508
	IC	5	2**	0			0 (0,20)	****	1.000	****	1.890*
	A. lyrata	26	14	3		86	5 (1,15)	*****	0.638	****	-0.271
HAT4	*A. lyrata ssp. lyrata*	8	5	0	**5.7**	8	0 (0,30)		0.300		-1.757*
	A. lyrata ssp. petraea	10	8	1		38	9 (1, 68)		0.412	*	0.989
	SC	6	6	1			25 (1,>500)		0.438		0.184
	IC	4	2	0			0 (0,31)	**	1.000	**	-0.840
	A. lyrata	18	12	3		34	25 (6, 70)		0.250		0.444
PGIC	*A. lyrata ssp. lyrata*	9	7	1	**5.0**	15	2 (1,15)		0.387		-0.615
	A. lyrata ssp. petraea	10	5***	0		67	0 (0,7)	*****	0.782	****	2.301*
	SC	4	4	0			0 (0,51)		0.788	*	-0.777
	IC	5	3	0			0 (0,130)	**	0.992	**	1.878*
	A. lyrata	18	11	3		58	3(1,14)	*****	0.250		1.029
SCADH	*A. lyrata ssp. lyrata*	12	4	0	**2.1**	5	0 (0,>500)		0.0		0.385
	A. lyrata ssp. petraea	14	7	2		22	5 (1,16)	**	0.392	*	-0.183
	SC	6	5	0			0(0,18)		0.438		1.15
	IC	8	3	0			0(0,13)	**	0.917	***	-1.78*
	A. lyrata	26	10*	2		20	2(1,10)	*****	0.377	**	0.0770

[a], estimate of recombination rate from *A. thaliana* mapping data

[b], expected value of the population recombination parameter ρbased on mapping data assuming the upper-bound mutation rate estimate, and r_{map}

[c], estimates of ρ from polymorphism data, using the method of Wall (2000). Approximate 95 % credibility intervals are shown in parentheses. Likelihoods are calculated over a grid of integer values for ρ.

[d], probability of the observed h and Rm or less likely configurations, assuming ρ=E(ρ). [e], probability of the observed B or greater, assuming ρ=E(ρ).

*, p<0.05, **, p<0.01 ***, p<10^{-3} ****, p<10^{-4} *****, p<10^{-5}

recombination during the history of the sample; R_m is greater than 0 for all loci except CAUL (Table 4.5); under the infinite sites model, this is only possible if a recombination event has occurred. Estimates of the physical recombination rates from the *A. thaliana* genome suggest that most of these loci are close to the average rate of crossing over in *A. thaliana* (r_{map}=5.0 cM/MB), which would predict high ρ values given the estimates of θ, even with the conservative upper-bound estimate of the mutation rate (Table 4.5).

All expected values of ρ using map-based estimates of recombination (E(ρ))are higher than estimates obtained from polymorphism data ($\hat{\rho}$), consistent with the hypothesis of excess linkage disequilibrium in the sampled loci in *A. lyrata ssp petraea* (Table 4.5, Figure 4.4), or a generally lower rate of crossing over in *A. lyrata* compared with *A. thaliana*. If we assume that the true recombination parameter is E(ρ), all loci except HAT4 show a significant departure from neutral predictions in *A. lyrata ssp. petraea* for the joint values of h and Rm, and all loci except CAUL show significance for B (Table 4.5, Figure 4.4). This is also evident within Scotland and Iceland, suggesting that the pattern is not simply the result of geographic structuring of haplotypes within this subspecies. No locus in *A. lyrata ssp. lyrata* departs significantly from neutral expectation assuming $4N_e r$=E(ρ), although the small number of segregating sites means that there is low power to detect departures from equilibrium (HUDSON 2001; WALL 2000), and the confidence intervals for $\hat{\rho}$ are often wide (Table 4.5, Figure 4.4). With few segregating sites, the wide confidence intervals imply that it is common for R_m to be close to or equal to 0 under the standard neutral model, even with a high ρ value.

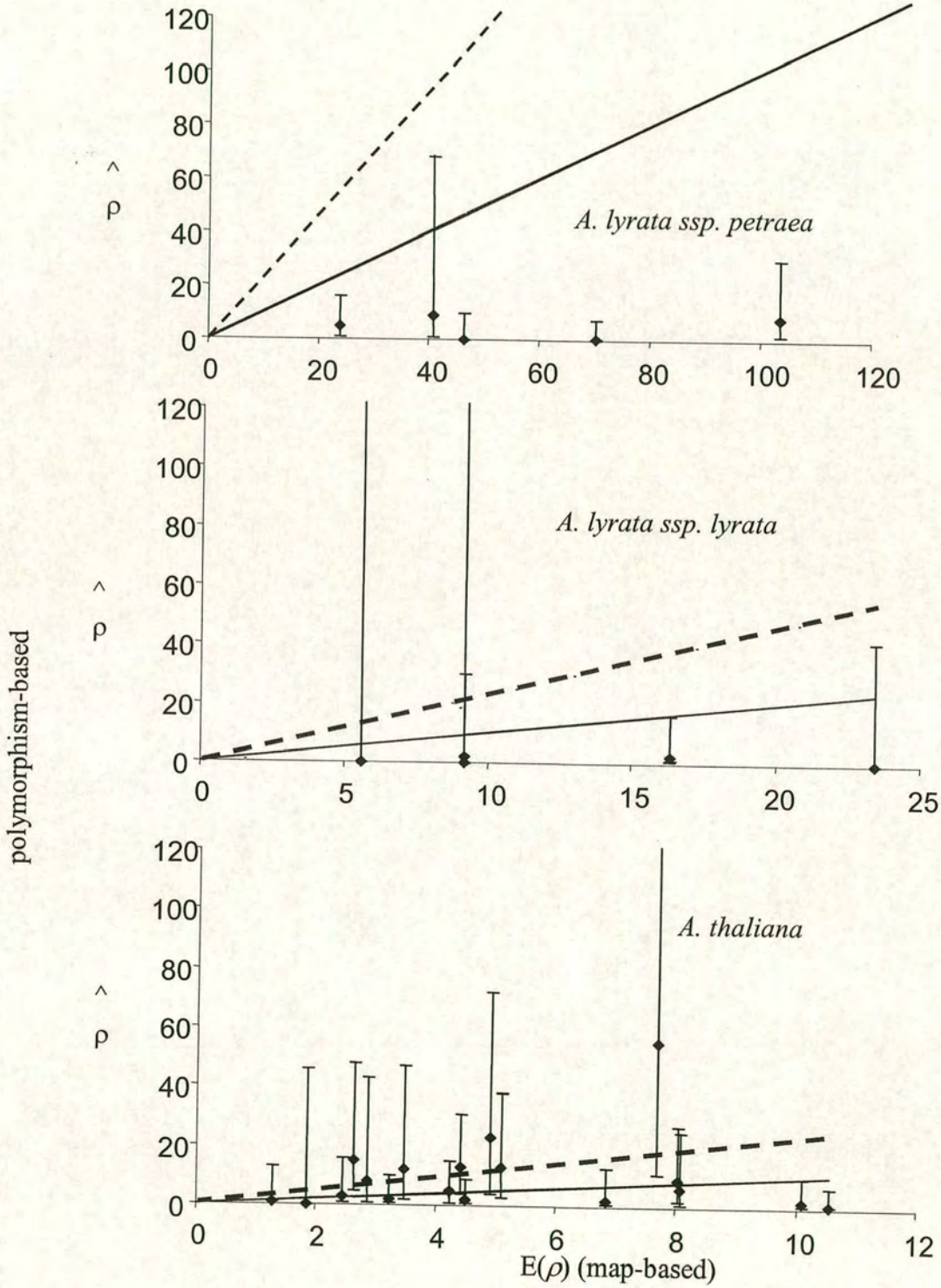


Figure 4.4 Relationship between observed and expected values of the per-locus recombination parameter ρ. 95 % credibility intervals from maximum likelihood estimates are given. Solid line shows the expected ρ assuming the upper-bound estimate of the mutation rate of 1.5×10^{-8}. Dashed line shows the expected values assuming the lower-bound mutation rate estimate of 6.5×10^{-9}.

In contrast with evidence for excess linkage disequilibrium in *A. lyrata*, estimates of $\hat{\rho}$ from worldwide samples of *A. thaliana* are very close to those expected (E(ρ)) under the standard neutral model, assuming 96% selfing (Figure 4.4). There is no consistent difference between map-based (E(ρ)) and polymorphism-based ($\hat{\rho}$) estimates of ρ, and few loci show any significant departure from expectation using either estimate of the mutation rate. This lack of significant departure for both u estimates suggests that the confidence intervals for $\hat{\rho}$ are wide in comparison with the uncertainty in mutation rates.

In *A. lyrata* Tajima's D statistic is significant under the assumption of no recombination for five individual tests performed (Table 4.5). Three cases (PGIC, ETR1 and SCADH) are for the Iceland population, and are due to haplotype structure. For example, ETR1 has two haplotypes in a sample size of five, generating a deficiency of low frequency polymorphisms, and a positive Tajima's D. Similarly, the SCADH locus has only two haplotypes in a sample size of 8 in Iceland, but there is a single representative of one haplotype generating a negative D. The negative Tajima's D for *A. lyrata ssp. lyrata* for HAT4 is generated by a single unusual haplotype in North Carolina, which is most similar to haplotypes found in Europe. Overall, the average Tajima's D is significantly greater than expected under the standard neutral model for *A. lyrata ssp. petraea*, but not significantly different from expectation for *A. lyrata ssp. lyrata* (Table 4.6). Within populations, the Iceland sample also shows an excess variance in Tajima's D. For *A. thaliana*, the average Tajima's D is significantly too low for the 18 loci included in the analysis, and the variance in Tajima's D is too high (Table 4.6).

Table 4.6-Multilocus test of Tajima's D[a]

population	Mean Tajima's D	% lower than mean	% higher than mean	Variance	% with lower variance	% with higher variance
A. lyrata ssp. lyrata	-0.26	33.4	66.6	1.54	88.7	11.3
NC	0.222	76.4	23.6	0.863	69.0	31.0
MI	-0.521	16.7	83.3	0.196	18.2	81.8
A. lyrata ssp. petraea	0.848	99.4	0.6	0.814	55.2	44.8
IC	0.201	73.3	26.7	2.73	99.98	0.02
SC	0.060	61.4	38.6	0.774	39.4	60.6
A. thaliana	-0.659	0.3	99.7	1.6	99.7	0.3

[a], percentages are the proportion of simulations that give lower and higher values for the mean and variance of D than observed for the nuclear genes

4.4 Discussion

4.4.1 Levels of polymorphism in A. lyrata

We observe high average estimates of within-population diversity in *A. lyrata*; the average of MLEs for θ within populations is 0.014, which is comparable to the average in *D. melanogaster* populations (MORIYAMA and POWELL 1996). Our estimates of total and within-population diversity are considerably higher than those found by Savolainen and colleagues at the ADH locus (SAVOLAINEN *et al.* 2000). This is probably because that study concentrated primarily on a population from the North American subspecies *lyrata*, which our results also show to have low polymorphism. Savolainen and colleagues report a π_{total} value of 0.0023 for this population, within the range of our estimates for *A. lyrata* ssp. *lyrata* (from 0.00098 for ETR1 to 0.0074 for PGIC). Estimates of polymorphism in the ADH locus suggested similar or lower levels of polymorphism in subspecies *petraea*, but this may simply reflect the high variance in parameter estimation associated with small sample sizes (Iceland: n=2 and Russia: n=4) (SAVOLAINEN *et al.* 2000), especially when there is strong haplotype structure, as found in our slightly larger samples. Our results are thus not incompatible with those for ADH, and the combined data

highlight the importance of surveying polymorphism for multiple loci, and including within- and between-population variability across the species range.

4.4.2 Comparing nucleotide diversity in selfing vs. outcrossing Arabidopsis

The average within-population diversity in our *A. lyrata* sample (θ=0.014) is much higher than estimates from within *A. thaliana* populations (θ=0.0013; Stahl *et al.* unpublished data). It is possible that this difference may be underestimated, since these *A. thaliana* populations were chosen because they were polymorphic at the disease resistance locus, *Rpm1*. On the other hand, they are all from North America, which *A. thaliana* is thought to have colonized recently. Nevertheless, the difference in within-population diversity suggests that the long-term effective size of outcrossing *A. lyrata* populations may be considerably higher than its selfing congener, and the magnitude of the difference suggests that increased homozygosity alone cannot explain the difference. However, there is also strong geographic structuring of polymorphism in *A. lyrata*, both within and between subspecies, suggesting a substantial amount of population isolation. North American populations have more than fivefold lower polymorphism than the Scotland and Iceland populations studied. Given the strong differences in the amount and types of segregating variants, the two subspecies should probably be considered independent taxa for studies of molecular population genetics. In *A. lyrata ssp lyrata*, the estimated average within-population silent polymorphism is only about twice that in *A. thaliana*, which is not significantly greater than the twofold expected under a neutral model with complete self-fertilization (Figure 4.3).

Species-wide diversity is also considerably higher in *A. thaliana* than *A. lyrata ssp. lyrata*, but lower than *A. lyrata ssp. petraea*. Furthermore, levels of population differentiation appear to be high in both species, although the highly polymorphic European *A. lyrata* populations show considerably less structure than *A. thaliana*, where polymorphism is strongly reduced within populations. Finally, patterns of linkage disequilibrium and the frequency spectrum in *A. lyrata ssp. petraea* suggest consistent, multilocus departures from the neutral equilibrium model. Together, the data suggest that diversity has been strongly influenced by population subdivision and demographic history, and estimates of θ and ρ should thus be treated with caution, and serve more as tests for neutrality and summaries of diversity patterns than as estimates of population parameters. Combined with similar evidence for departures from demographic equilibrium in *A. thaliana*, this makes quantitative comparisons of the long-term effective population sizes of these taxa difficult, since they depend on unknown details of the species' demographic history.

4.4.3 Demographic history in A. lyrata

The difference in polymorphism levels between subspecies in *A. lyrata* is consistent with several possibilities. First, polymorphism in the North American populations may have been reduced by a strong population bottleneck. This may have been a recent event associated with glaciation, and postglacial re-colonization, or it could have occurred during an initial colonization of North America from Europe, followed by a long-term reduction in effective population size. A recent bottleneck should generate a skewed frequency spectrum of variants, detectable by Tajima's D (TAJIMA 1989a). In the first stages after a bottleneck, a positive D is

expected, because intermediate frequency variants are most likely to be sampled from the ancestral population. Following this period, new mutations are expected to accumulate in different lineages of the expanding population, leading to a star-like phylogeny and a negative Tajima's D (FAY and WU 1999; TAJIMA 1989a). Savolainen and colleagues (2000) found a significant positive Tajima's D in a large sample from an Indiana population of *A. lyrata ssp. lyrata*, suggesting a population bottleneck. In our pooled sample from two subpopulations of *A. lyrata ssp. lyrata*, the mean and variance of Tajima's D are close to their neutral expectation (Table 4.6). However, given preliminary evidence for population subdivision between Michigan and North Carolina, the signature of a bottleneck may be obscured due to geographic structuring of variation. Larger within-population samples will be required to further investigate the frequency spectrum of neutral polymorphisms within North America, to determine whether these populations are at equilibrium with respect to mutation and genetic drift.

An alternative to the bottleneck hypothesis is that the ssp. *lyrata* diversity levels represent the long-term equilibrium situation, and that recent population admixture or introgression from previously isolated populations has inflated polymorphism in the European populations. Coalescent models applied to human population data suggest that, to be detectable, admixture must be very recent, and must involve highly diverged populations (NORDBORG 2000b). For *A. lyrata*, population admixture between populations from distinct glacial refugia (see e.g. WALTER and EPPERSON 2001) is one possibility, although it is questionable whether divergence during the last glaciation could have been long enough to cause the observed level of differentiation among haplotypes. Another possible source of

excess polymorphism could be recent introgression from related species, for example *A. halleri*, which has a patchy distribution in Western Europe (JONES, 1993). A recent study of species-wide polymorphism at several loci in *A. lyrata* and *A. halleri* supports this possibility for some *A. lyrata* populations (B. Stranger, pers. comm.), but it is unclear whether it can account for the results from Iceland and Scotland, where *A. halleri* is absent (JONES, 1993); further work is necessary to test whether variants in these populations are shared with *A. halleri*.

The recent admixture hypothesis predicts an excess of linkage disequilibrium. Our evidence for strong haplotype structure using both within-population samples and the pooled sample in Scotland and Iceland may thus be consistent with this hypothesis; if so, effective population sizes based on levels of nucleotide diversity could be considerably overestimated. The signature of excess linkage disequilibrium is most striking in Iceland, where most loci have very few haplotypes, despite extensive polymorphism, and the average Tajima's D is positive. Alternatively, the excess linkage disequilibrium and positive Tajima's D values could result from a population bottleneck in Europe, with an ancestral population with a high effective size and high diversity. It is also possible that long-term population subdivision with low levels of migration could generate the observed increase in linkage disequilibrium and skew in the frequency distribution (Wakeley and Aliacar, 2001). Analysis of haplotype structure and the frequency spectrum using larger samples within populations will be important to confirm these results, and to test further for departures from equilibrium assumptions within populations.

4.4.4 Recombination and linkage disequilibrium in A. lyrata and A. thaliana

Our results suggest strong linkage disequilibrium in *A. lyrata* compared to the amount expected based on the r_{map} values from *A. thaliana*. The available polymorphism datasets thus provide little evidence for the expected higher effective recombination rate in *A. lyrata* compared to *A. thaliana*. In *A. thaliana*, the effective intralocus recombination rate in worldwide samples of alleles appears to be approximately as predicted from the estimated recombination and mutation rates, given this species' high rate of self-fertilization. Our result using mutation and recombination parameter estimates corroborates the conclusions of Nordborg and colleagues (HAGENBLAD and NORDBORG 2002; NORDBORG 2000a) which were based on comparing their estimates of linkage disequilibrium from several large genomic regions in *A. thaliana* with those observed in *D. melanogaster*. Together, the results do not support previous claims of a general excess of recombination in *A. thaliana* compared with the amount expected in a highly selfing organism (HAUBOLD *et al.* 2002; KUITTINEN and AGUADE 2000), and so a high rate of gene conversion (HAUBOLD *et al.* 2002) may not be necessary to explain the observed number of recombination events in this species. If the general assumption of a high long-term selfing rate is correct, and provided that rates of gene conversion are not excessive, this good fit with neutral expectation also suggests that there is no strong population subdivision in these worldwide samples of *A. thaliana*, in contrast with earlier suggestions (INNAN *et al.* 1996). However, if there has been recent population expansion, the signature of ancient population subdivision may be obscured in *A. thaliana*. Population growth leads to a star phylogeny with few significant associations among the polymorphic sites due to variants that have arisen during the

growth period. Growth thus increases the number of rare variants and the number of haplotypes. If a population has experienced growth long enough ago to accumulate new variants, linkage disequilibrium will therefore be low, as measured by haplotype number or the squared correlation coefficient r^2 (PRITCHARD and PRZEWORSKI 2001; PRZEWORSKI and WALL 2001; SLATKIN 1994). This could obscure linkage disequilibrium caused by long-term subdivision, and generate a misleading appearance of fitting the expectations of recombination-drift equilibrium.

Evidence in *A. lyrata* for consistently lower than expected estimates of ρ is similar to observations from some population samples of Drosophila (ANDOLFATTO and PRZEWORSKI 2000; WALL *et al.* 2002) and humans (FRISSE *et al.* 2001). In all these species, population bottlenecks following colonization from ancestral populations may explain the results. An important limitation of studies of linkage disequilibrium in *A. lyrata*, however, is the absence of any extensive information on rates of recombination for this species. One alternative to a demographic explanation is the possibility that in *A. lyrata* the loci surveyed are in regions of reduced recombination. Another possibility is that recombination rates are generally lower in *A. lyrata* than *A. thaliana*. Data on chiasma frequencies have suggested that rates of recombination may be generally higher in selfing species than related outcrossers, and this accords with some models for the evolution of recombination (reviewed in (CHARLESWORTH and CHARLESWORTH 1979). The average r_{map} in *A. thaliana* is approximately five times higher than rates in regions of normal recombination in humans and *D. melanogaster*, suggesting the possibility of elevated recombination rates in *A. thaliana*. It is not known how large the effect, if any, may be, but we can roughly calculate the magnitude of the difference in recombination rates in the two

Arabidopsis species that would be required to explain our data. The average estimate of $\hat{\rho}/\hat{\theta}$ from our *A. lyrata ssp. petraea* datasets is 0.42. Under the standard neutral model, this gives an estimate of the ratio of the rate of recombination to the rate of mutation. A mutation rate of 1.5×10^{-8} would then imply that the average *r* for these loci is 0.6 cM/MB in *A. lyrata*, approximately 7 times lower than the average of 4.6 cM/MB estimated from physical and genetic maps for these loci in *A. thaliana*; if the mutation rate is closer to the estimate of Wolfe et al. (1987), this would suggest a 13-fold reduction in recombination rates in *A. lyrata*. It would be surprising if these species differ so greatly in recombination rates, but empirical data are needed to test this. For comparison, the average $\hat{\rho}/\hat{\theta}$ for *A. thaliana* is 0.932, greater than twice that of *A. lyrata ssp. petraea*. Assuming a selfing rate of 0.96, the average recombination rates are estimated as 5.6 or 14.8 cM/MB, using the lower-and upper-bound mutation rate estimates, respectively. These values are of the same order as the average of 5.1 cM/MB estimated from mapping data.

Chapter 5

Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*.

5.1 Introduction

Transposable elements (TEs) in many organisms have been shown to accumulate differentially among chromosomal regions, including regions of contrasting recombination rates (BARTOLOME *et al.* 2002; BOISSINOT *et al.* 2001; CHARLESWORTH and LANGLEY 1989; DURET *et al.* 2000), gene density (MEDSTRAND *et al.* 2002), and base composition (LANDER *et al.* 2001). One possible explanation for these patterns is that TEs have insertion preferences for particular regions. Evidence for insertion bias is strong for some TEs (JAKUBCZAK *et al.* 1991), although for the majority there is little evidence to support insertion preference as an explanation for TE distribution (NUZHDIN *et al.* 1997). An alternative is the effects of differential selective constraints on TEs in different regions of the genome (CHARLESWORTH and LANGLEY 1989). First, TE's are almost exclusively found outside of coding regions (BARTOLOME *et al.* 2002; DURET *et al.* 2000; NEKRUTENKO 2001; PAVLICEK *et al.* 2002), suggesting that the majority of insertions into exons are strongly deleterious and rapidly eliminated by selection. Additionally, evidence from patterns of TE insertion polymorphism in natural populations of some species indicate that insertions segregating in noncoding genomic regions are almost always at low frequencies, consistent with the hypothesis that TE abundance is controlled by the action of purifying selection (CHARLESWORTH and LANGLEY 1989; WRIGHT *et al.* 2001). If TE abundance in

noncoding DNA is determined by a balance between the forces of transposition and natural selection, regional genome effects on the strength or efficacy of natural selection will play a significant role in controlling TE distribution.

Several models of selection against TEs in noncoding DNA have been proposed. First, abundance may be controlled by weak selection against the direct effects of insertions into noncoding regions (BIEMONT *et al.* 1997; CHARLESWORTH and LANGLEY 1989). In particular, insertions into introns and regulatory regions may have on average slightly deleterious consequences on fitness (LANDER *et al.* 2001; LONG *et al.* 2000) by causing disruptions in gene expression. Similarly, element transposition may impose a significant cost on the host due to expression of TE gene products, leading to selection against TE activity (NUZHDIN *et al.* 1996). Alternatively, the action of natural selection may be only indirectly associated with TE mobility, by the deleterious effects of ectopic recombination between elements located at distinct sites in the genome, which can cause major chromosomal rearrangements and gene deletions (LANGLEY *et al.* 1988).

Under the ectopic exchange model, lower rates of ectopic exchange are expected in regions of reduced recombination (VIRGIN and BAILEY 1998), allowing TEs to accumulate in these chromosomal locations (LANGLEY *et al.* 1988). However, under the insertion model, the action of positive and negative selection at linked sites may also weaken the efficacy of selection against deleterious insertions in regions of reduced recombination (DURET *et al.* 2000; EICKBUSH and FURANO 2002), a process known as the Hill-Robertson effect (HILL and ROBERTSON 1966). Although further modelling is required to assess the action of Hill-Robertson interference on TEs, the effects of linked selection may thus also allow elements to

accumulate in regions of reduced recombination. Unlike the ectopic exchange model, however, the insertion model predicts that TE insertions should accumulate in regions of low gene density; even within noncoding DNA, TE insertions are less likely to interfere with gene expression in regions with a low proportion of coding DNA.

Results showing higher copy numbers of TEs in regions of reduced recombination in several species (Arabidopsis Genome Initiative 2000; BARTOLOME *et al.* 2002; BOISSINOT *et al.* 2001), are consistent with this expectation of both models. Although most of these studies have examined the effects of large-scale heterogeneity in recombination, a recent genetic analysis in maize also provided evidence for a strong reduction in the rate of recombination in TE-rich regions at a very local level (FU *et al.* 2002). Furthermore, a recent study of TE frequencies in populations of *Drosophila melanogaster* has found evidence for an effect of TE size on the action of selection, which was interpreted as additional support for the model of ectopic exchange (PETROV et al. 2003). In contrast with these results, however, a recent study showed a positive correlation between recombination rate and DNA transposon abundance in *C. elegans*, and no effect of recombination on retrotransposon abundance (DURET *et al.* 2000), suggesting that the predictions of all selection models do not hold. Although the explanation for the differences among species is unclear, *C. elegans* is highly inbreeding in natural populations (GRAUSTEIN *et al.* 2002). In highly inbred species, low effective rates of recombination across the genome may remove effects of recombination rate heterogeneity on genome structure (CHARLESWORTH and WRIGHT 2001; MORGAN 2001). First, unless population size or physical recombination rates are very high,

the action of Hill-Robertson interference is expected to be present genome-wide (McVean and Charlesworth 2000), and any recombination-based heterogeneity associated with the efficacy of selection may be weak or absent. In addition, high homozygosity in self-fertilizing populations is expected to lead to infrequent ectopic recombination events, due to fewer chances for ectopic pairing, which appears to be promoted by heterozygosity (Charlesworth and Charlesworth 1995; Morgan 2001; Wright *et al.* 2001). Patterns of heterogeneity in TE distributions observed in the genomes of highly inbred species should thus reflect other effects of natural selection, or the effects of transposition preferences, and variation in recombination rate is less likely to play a role.

To further evaluate this hypothesis, we investigate the effects of recombination rate and gene density on TE distributions in the self-fertilizing plant *Arabidopsis thaliana*. *A. thaliana* shows a high diversity of TEs, most of which are widely dispersed across the genome (Le *et al.* 2000; Surzycki and Belknap 1999). This diversity includes all major superfamilies of retrotransposons (Class I elements), DNA transposons (Class II elements), as well as a recently identified third class of TEs ('Class III') (Kapitonov and Jurka 2001; Le *et al.* 2000), with similarities to bacterial rolling-circle transposons (Kapitonov and Jurka 2001). Genome analysis has provided preliminary evidence for an accumulation of many TE families in pericentromeric regions (Arabidopsis Genome Initiative 2000; Copenhaver *et al.* 1999; Kumekawa *et al.* 2000), but the relative importance of various characteristics of genome structure in driving this pattern has not been investigated.

5.2 Methods

Genome analysis

The genome sequence and positions of predicted genes for all five chromosomes of *A. thaliana* were extracted from The Arabidopsis Information Resource (TAIR- www.arabidopsis.org). The genome sequence is in the form of 'pseudomolecules'; the concatenated sequence of overlapping sequenced clones for each chromosome. These pseudomolecules comprise the entire genome sequence, with the exception of telomeres, the rDNA islands on chromosomes 2 and 4, and several small gaps denoted as 'N's (Arabidopsis Genome Initiative 2000). Sequences were divided into 100kb fragments across each chromosome, and information on base composition and annotated (predicted) gene position was collected, excluding a small number of fragments with estimated gaps or terminal sequence leading to a bin size of less than 50 kb. Each pseudomolecule was submitted to a BLAST search (NCBI standalone BLAST released for Unix on April 3,2001) against our Arabidopsis TE database (www.tebureau.mcgill.ca/), to determine the positions of each TE superfamily along the entire set of chromosomes. TE positions were then verified manually, to eliminate all redundancies and record the presence of nested insertions. For analysis of gene density and the fraction of coding DNA, all annotated coding regions that showed matches solely to TE–encoded gene products were removed. All putative insertions into annotated exons and introns were verified by individual BLAST searches of annotated coding regions. Matches to small internal fragments of TEs, annotated proteins that represent fusions between TE mobility genes and adjacent genes, and matches to host sequences internal to TE insertions were eliminated.

Recombination rate estimation

The alignment of physical and genetic maps from *A. thaliana*, and estimation of recombination rates across the genome were as described in Chapter 4. Boundary regions with low marker density between the central regions of low recombination and the chromosome arms were excluded from correlations with recombination rate.

Central regions of reduced recombination, which include the centromeres and heterochromatic knobs, were also delimited using the genetic marker data in order to compare these regions with the rest of the genome, hereafter called the 'chromosome arms'. The central regions have been shown to have a strong reduction in rates of recombination, by recombination rate estimation that is independent of the genome project data used here (COPENHAVER *et al.* 1999; HAUPT *et al.* 2001). Because of low marker density, comparisons were restricted to central regions of clearly reduced recombination versus the chromosome arms, with boundary regions of low marker density excluded from the analysis. In the case of chromosome 1, fine-scale estimates of recombination rates in the centromeric region by Haupt et al. (2001) allowed the recombination boundary on the short arm to be precisely delimited, to BAC clone T22A15. Estimates of average recombination rate along the arms (4.9 cM/MB) using the polynomial fits were very similar to estimates of the average genome recombination rate, 5.0 cM/MB (HAUPT *et al.* 2001). In the central regions of reduced recombination, average recombination rate using polynomial fits was estimated to be 1.5 cM/MB, although this includes regions with much greater recombination rate reduction, up to perhaps complete suppression in the centromere core (HAUPT *et al.* 2001).

Testing effects of coding density on purifying selection against insertions.

We used a maximum likelihood method to test for the presence of selective constraint in intergenic regions. If TE insertions are inserted randomly within intergenic DNA along the chromosome arms, they should follow a Poisson distribution. Specifically, the likelihood of observing n insertions is the product of the likelihoods in each bin i:

$$L = \prod L_i = \prod \frac{u_i^{n_i}}{n_i! e^{u_i}}$$

Assuming no selective constraint in intergenic regions, the parametric mean $\mu_i = (\nu I_i)$, where I_i is the number of base pairs of intergenic DNA in bin i estimated directly from the data, and ν is the per base pair probability of an insertion. This model assumes complete independence among bins, i.e. no local insertion preference, and by constraining ν across bins we assume that TE distribution is random across intergenic DNA. We can calculate the maximum likelihood estimate of ν, using the observed values of n and I for each bin. As an alternative model, we allow u_i to vary as a function of the observed amount of coding DNA (C_i) in bin i: $u_i = \nu(I_i - \alpha C_i)$, where $\alpha C_i < I_i$. Here, $\frac{\alpha C_i}{I_i}$ is a measure of the proportion of intergenic DNA in bin i that is under selection against TE insertion, and the total proportion of intergenic DNA under selective constraint against TE insertions can be estimated by summing across all bins. We assume that a given insertion site in intergenic DNA is either unconditionally deleterious, with probability $\frac{\alpha C_i}{I_i}$, or

neutral, with probability $1-\frac{\alpha C_i}{I_i}$. By calculating the joint maximum likelihood of v and α, and comparing it to a model where $\alpha=0$, we can test for an effect of coding density on TE distributions. Significant improvement to the likelihood is assessed using the statistic $2(\ln L_1 - \ln L_0)$, where L_0 assumes $\alpha=0$, and L_1 allows $\alpha>0$. This statistic is asymptotically χ^2 distributed with 1 degree of freedom. To examine the expected correlation between coding density and TE frequency under a model with these parameter estimates, we also run 10000 computer simulations of random TE insertion into unconstrained intergenic DNA, using the observed sizes of intergenic and coding DNA in each bin. In these simulations, we simply assume a Poisson distributed number of insertions into each bin, with the parametric mean u_i in bin i given by the selection equation above. These simulations assume that the only factor influencing TE distribution is random insertion into unconstrained sites.

5.3 Results

5.3.1 TE accumulation in low-recombining regions surrounding the centromere

Central regions of reduced recombination, including the centromeres, pericentromeric regions and heterochromatic knobs, have been shown to have a strong reduction in rates of recombination (COPENHAVER *et al.* 1999; HAUPT *et al.* 2001), and we first compare TE abundance in these regions with the chromosome arms. All classes of TEs show evidence for accumulation in these regions of reduced recombination surrounding the centromere (Arabidopsis Genome Initiative 2000). In total, TE-derived DNA represents approximately 27% of the central regions of

reduced recombination, compared with only 3% of DNA in the rest of the genome. Two superfamilies of elements, *gypsy*-like and CACTA, show a particularly striking accumulation in the centromeric regions, and are almost exclusively found in these regions, and we therefore consider these elements separately. These TEs, particularly the *gypsy-like* element *Athila*, have many centromeric copies arranged in tandem, often as truncated or fragmented elements, located near the centromeric core (KUMEKAWA *et al.* 2000; PELISSIER *et al.* 1995). Such elements have been hypothesized to be associated with centromere function (MALIK and HENIKOFF 2002), and they may have acquired new mutational mechanisms for their spread in the centromere. However, all other elements, which are not found as tandem arrays, also show significant accumulation in pericentromeric regions of reduced recombination.

In addition to low levels of crossing-over, *A. thaliana* pericentromeric regions also have low exon density, and the fraction of DNA that is coding is much lower compared with the rest of the genome (Table 5.1, see Arabidopsis Genome Initiative 2000). Similar to recent studies in *Drosophila melanogaster* (COMERON and KREITMAN 2000), average intron length is significantly larger in the regions of reduced recombination, even when the contribution of TE insertions to intron size is factored out (Table 5.1). Overall differences in base composition are also significant, but the difference in mean base composition between regions is very weak. Surprisingly, GC content is slightly higher in the pericentromeric regions compared with the chromosome arms (Table 5.1). This contrasts with the pattern at the local

Table 5.1 TE copy number (per MB) and genome characteristics in regions of reduced recombination compared with the chromosome arms.

	reduced recombination	chromosome arms
Class I[1]	16.3***	3.5
gypsy	33.0***	0.3
Class II	30.6***	10.2
CACTA	6.5***	0.16
Class III	16.9***	6.7
GC content	0.368**	0.361
Exon density (per Mb)	417**	1300
Fraction of coding DNA	0.131***	0.326
Intron size (bp)[2]	188.9**	167.2

[1], Class I elements include those insertions that could be definitively classified as *copia*-like, LINE-like, and SINE-like
[2], excludes contribution of TE insertions to intron length
Significance levels are given for the Mann-Whitney U test, *, $p<0.05$; **, $p<0.01$; ***; $p<0.001$

level, where base composition surrounding TE insertions has been found to show an elevated frequency of AT (LE *et al.* 2000).

Because gene density is higher in regions of normal recombination, one simple explanation for the differential accumulation of TEs in low-recombining regions is that it reflects the action of strong purifying selection against insertions into exons. Indeed, we find an almost complete absence of insertions into exons (Table 5.2), suggestive of very strong selection. Note that we exclude from this category 1) annotated genes matching solely to TE-derived gene products, 2) 'fusion' proteins, encoding predicted genes which fuse coding sequence from TE insertions and adjacent genes, and 3) 'acquired genes' that are host genes internal to TE insertions (YU *et al.* 2000). Although both fusion proteins and acquired genes may be expressed and functional, they do not represent the insertion of TEs into pre-existing host coding regions. TE insertions are also strongly under-represented in introns in *A. thaliana* (Table 5.2); in both the centromere and the chromosome arms,

there is a strong reduction in the frequency of TE insertions in introns. Since most insertions into introns and exons appear to be under

Table 5.2- Location of TE insertions in relation to coding regions[1].

Element Class	Reduced Recombination			Chromosome arms		
	intergenic	intron	coding	intergenic	intron	coding
Class I[2]	19.4*** (0.049)	2.99* (0.0004)	1.4* (0.0007)	5.8 (0.014)	0.378 (0.0002)	0.07 (0.0002)
Class II[2]	35.5*** (0.046)	9.4*** (0.023)	0	18.1 (0.018)	1.5 (0.0001)	0
Class III	20.9*** (0.014)	3.2 (0.002)	0.56 (0.0005)	12.5 (0.006)	1.1 ($9.8x10^{-4}$)	0

[1], Upper values in each cell are the number of elements per megabase of DNA in each category, and values in parentheses are the fraction of DNA in each location occupied by TEs. Significance levels are shown for comparisons between the regions of reduced recombination compared to the chromosome arms by the Mann-Whitney U test.
[2], Class I and Class II elements are shown excluding *gypsy* and CACTA-like elements, respectively.
two-tailed significance: *,$p<0.05$; **, $p<0.01$, ***$p<0.001$

strong purifying selection, this might be the sole explanation for their low abundance in the chromosome arms, where the density of coding regions is higher. When the abundance of TEs within intergenic DNA is compared, however, TEs are still significantly in excess in regions of low recombination (Table 5.2), suggesting that the difference in the abundance of coding DNA alone cannot explain their differential accumulation.

5.3.2 Recombination rate does not explain TE distribution in intergenic regions of A. thaliana

The high number of TEs in pericentromeric regions is consistent with either model of selection, since these regions are both gene-poor and recombination-poor. Furthermore, the dense methylation and difference in chromatin structure in

heterochromatic regions may make these insertions more effectively silenced (SINGER *et al.* 2001), preventing the potentially deleterious expression of TE-derived gene products in these regions. It is therefore important to examine whether, in addition to this comparison between genomic regions, there are general correlations between TE abundance and recombination rate. Given the under-representation of TEs in coding regions, we consider only the TE content in intergenic regions for this analysis. Excluding the centromere-specific TEs (*gypsy* and CACTA), no correlation is observed between the rate of recombination and TE abundance in intergenic DNA (Table 5.3). Note that this whole-genome correlation with recombination is not significant despite accumulation of elements near the centromere; since the pericentromeric regions represent a low proportion of total DNA, even this effect is not detected in a genome-wide correlation. Only *gypsy* and CACTA, which are almost exclusively present in the central regions of reduced recombination, show a significant negative correlation. If we consider only the chromosome arms, no negative correlation is observed for any class of TE; in fact, all three major classes show a positive relationship with TE frequency. This positive relationship can be explained at least in part by a negative correlation between coding density and recombination rate along the chromosome arms (Table 5.3). Using a partial correlation correcting for the effect of coding density, there is no significant effect of recombination rate on Class II or III elements, and we observe

Table 5.3 Spearman rank correlation coefficients between element frequency, fraction of coding DNA and recombination rate.[1]

	total genome		chromosome arms	
	recombination rate	fraction of coding DNA[2]	recombination rate	fraction of coding DNA[2]
Class I	**-0.005** (0.067*)	**-0.375***** (-0.322***)	**0.127**** (0.0898*)	**-0.187***** (-0.202***)
Gypsy	**-0.378***** (-0.257***)	**-0.426***** (-0.469***)	**0.060** (0.124***)	**-0.057** (-0.117**)
Class II	**-0.011** (-0.006)	**-0.318***** (-0.232***)	**0.080*** (0.021)	**-0.100***** (-0.155***)
CACTA	**-0.247***** (-0.193***)	**-0.271***** (-0.276***)	**0.023** (0.0011)	**-0.074*** (-0.057)
Class III	**0.035** (0.09**)	**-0.244***** (-0.218***)	**0.076*** (0.037)	**-0.064** (-0.164***)
fraction of coding DNA	0.099*	-	-0.139**	-

[1]- values given in parentheses are partial correlations, factoring out effects of coding fraction and recombination rate, respectively.
[2]- Direct correlations between coding density and TE abundance are calculated by making bins which exclude TE contributions to DNA length. Partial correlations, factoring out recombination rate effects, use the original bins, for accurate estimation of recombination rates over real physical distance.
two-tailed significance: *,$p<0.05$; **, $p<0.01$; ***,$p<0.001$

a weaker positive relationship between recombination rate and TE abundance for Class I TEs (Table 5.3). Provided that the rate of ectopic recombination correlates with the rate of crossing over, these results suggest that selection against ectopic exchange does not drive TE distribution in *A. thaliana*.

5.3.3 Effect of gene density on TE abundance in intergenic DNA

Our initial analysis suggested that selection against insertions into exons and introns is playing an important role in TE distribution in the *A. thaliana* genome, given their strong under-representation compared with intergenic DNA. We also wish to test whether regions of high gene density show an under-representation of insertions in intergenic regions, due to a higher density of untranslated and *cis*-

regulatory regions. Alternatively, silencing of TE insertions by DNA methylation and associated changes in chromatin structure might interfere with the expression of nearby genes (GENDREL *et al.* 2002). In a given sample of DNA along the chromosome, increases in TE-derived DNA could force the amount of coding sequence to be smaller, violating the standard assumptions of correlation analysis. To avoid this effect, we excluded the contribution of TEs to sequence length when sampling bins of constant size (100kb) across the genome, to look at effects of coding density on TE frequency. In contrast with recombination rate, all element families show a significant negative correlation between the fraction of coding DNA and TE frequency in intergenic regions (Table 5.3). If we consider only the chromosome arms, the negative correlation remains significant for Class I and II elements, but not Class III (Table 5.3). Furthermore, a partial correlation analysis, factoring out the effects of recombination rate, still reveals a residual negative correlation between coding density and TE abundance for both the whole genome and the chromosome arms (Table 5.3).

While the above analysis suggests that selection against insertions near coding regions is influencing TE distribution, the corresponding correlation coefficients are low, and it is unclear how strong the action of purifying selection would be to explain the data. To assess the insertion model more directly, we use a maximum likelihood framework to test whether a simple model that includes an effect of coding density on the amount of selective constraint in intergenic regions provides a significant improvement to the likelihood of the data compared to a model that assumes elements are randomly distributed in intergenic regions (see Methods). Because the high fraction of TE-derived DNA and the high frequency of

nested insertions in the centromere will violate the assumptions of independence among insertions, we consider only the chromosome arms for this analysis, and once again exclude TE contributions to sequence length.

For all three TE classes, a model that includes increases in selective constraint in intergenic regions with an increasing amount of coding DNA provides a significant improvement to the likelihood compared to a model of random distribution in intergenic regions, consistent with the action of purifying selection against insertions near genes (Table 5.4). We can use this model to estimate α, a measure of the amount to which increases in coding DNA increase the selective constraint in intergenic regions. Estimates of α range from 0.27 for Class II elements to 0.64 for Class I. Additional factors, such as variation across individual element families and genomic regions in α, the presence of nested insertions, and the influence of historical selection pressures predating the evolution of self-fertilization may all play a role in causing residual departures from expectation. Nevertheless, we can use our analysis to roughly estimate the total fraction of intergenic DNA that is under purifying selection against TE insertions along the chromosome arms (see Methods). Using the total amounts of intergenic and coding DNA in our sample from the chromosome arms, our estimates of α would imply that between 18 and 44% of intergenic DNA is under selective constraint against TE insertions. Similarly, given an estimated average of 1308 bp of coding DNA per gene along the chromosome arms, this implies that each gene has on average 235-837 base pairs of intergenic DNA that are under selective constraint

Table 5.4- Likelihood ratio test for selection on insertions in intergenic DNA

Element Class	$\hat{\alpha}$ (95% C.I.)[1]	2 ΔL	r_s, model (95% C.I.)[2]
Class I	0.64 (0.52, 0.70)	31.1***	-0.178 (-0.124,-0.232)
Class II	0.27 (0.04,0.42)	4.9*	-0.07 (-0.01,-0.126)
Class III	0.34 (0.1, 0.49)	6.7**	-0.085 (-0.027, -0.144)

[1], maximum likelihood estimate and approximate 95% credibility interval for the estimate of α using the χ^2 approximation.
[2], mean and 95% confidence interval of the Spearman rank correlation coefficient between TE frequency and coding density from 10,000 computer simulations of random insertion, using the maximum likelihood estimates of α and v.
significance of likelihood ratio test: *,$p<0.05$; **, $p<0.01$; ***,$p<0.001$

against TE insertions. Given the short distance between genes in the compact *A. thaliana* genome (average length of intergenic DNA, 2013 bp), this conclusion appears reasonable. Computer simulations of insertion into unconstrained regions using the observed values of the abundance of coding and intergenic DNA and the maximum likelihood estimates of parameters confirm that the observed correlation coefficients between TE frequency and coding density are expected with this level of selective constraint (Table 5.4). This illustrates that even fairly strong selection against insertions in intergenic DNA will generate similar correlations between coding density and TE frequency along the chromosome arms to those observed.

5.3.4 Multivariate analysis of TE distribution

It is important to assess whether our analyses based on pairwise correlations and the three major TE Class categories are robust to a more global, multivariate analysis. To investigate this we used principal components analysis to summarize multivariate patterns of TE abundance in intergenic DNA for all individual TE element families, excluding CACTA and *gypsy*-like elements. The first axis in this

data reduction, which explains 23% of the variance, shows a general increase with TE abundance; all TE families increase in abundance with an increasing value for factor 1 (Table 5.5). We performed a multiple linear regression, using factor 1 as the dependent variable, with coding density, recombination rate, and GC content as independent variables. The model, which explains 20% of the overall variance in global TE abundance, provides confirmation of our primary conclusions. As expected, this analysis shows no significant effect of recombination rate on this

Table 5.5 principal component analysis of TE abundance. Values shown are correlations between element abundance and Factor 1.

TE family	correlation with factor 1
Ac	0.446
Basho	0.575
LINE	0.541
SINE	0.348
MULE	0.579
COPIA	0.442
MARINER	0.461
MITE	0.344

multivariate measure of TE abundance ($p>>0.05$), while there is a highly significant negative effect of coding density (regression coefficient=-0.42, $p<<0.05$). The model also shows a significant positive effect of base composition (regression coefficient=0.40, $p<<0.05$), presumably as a result of the increased GC content close to the centromeres (Table 5.1).

5.4 Discussion

In the genome of the self-fertilizing *A. thaliana*, there are strong effects of chromosome position on TE abundance, with a large accumulation of TEs close to the centromere. However, while there is a global difference between regions of contrasting recombination, our results suggest that recombination rate is relatively

unimportant in causing this pattern, and the effects of selection against TE disruption of gene expression may be the primary mode of selection affecting TE distribution in this self-fertilising species. Our results suggesting a low rate of ectopic recombination in *A. thaliana* are in accordance with a recent analysis of long terminal repeat (LTR) retrotransposons in this species (DEVOS *et al.* 2002), which indicated that small deletions within elements occurred at a much higher rate than ectopic recombination events between LTRs.

Along the chromosome arms, most TE families show slight positive correlations with recombination rate. This is similar to observations for DNA transposons in the holocentric genome of the self-fertilizing nematode *C. elegans*, where a general positive correlation between recombination rate and TE abundance is observed (DURET *et al.* 2000), while no effect of recombination is seen for retrotransposons. This correlation may result from the effects of biased transposition (DURET *et al.* 2000). However, a negative correlation between recombination and gene density in the euchromatic portions of *A. thaliana*, and a generally higher gene density in regions of low recombination in *C. elegans* (BARNES *et al.* 1995; WILSON 1999) suggest that these effects may at least in part also be driven by selection against insertions near genes.

Under the deleterious insertion model, higher gene density is expected to reduce the abundance of TEs in intergenic regions, due to a higher fraction of regulatory regions. However, our results (Table 5.2) and those in *C. elegans* (DURET *et al.* 2000) suggest that there is also an increase in TE abundance in introns in regions of low gene density. Nevertheless, genome analysis in both *A. thaliana* and *C. elegans* has provided evidence for a higher proportion of pseudogenes in regions

of low gene density (COPENHAVER *et al.* 1999; HARRISON *et al.* 2001) and a lower proportion of annotated genes in these regions are known to be expressed (C. elegans sequencing Consortium 1998; COPENHAVER *et al.* 1999), suggesting that this accumulation may also reflect a relaxation of natural selection against the disruption of gene function. Together, the results from both species are consistent with the hypothesis that mating system plays an important role in the evolution of genome structure.

It is important to consider whether the results could be consistent with alternative selection models to the deleterious effects of gene disruption. First, the results might also be interpreted as consistent with a model of selection against the deleterious expression of TE-derived gene products. In particular, TE insertions in regions of high gene density may show higher levels of TE-derived gene expression due to differences in chromatin structure, and may consequently impose stronger deleterious effects on host fitness. However, a large fraction of TEs in the Arabidopsis genome lack coding capacity, and many appear capable of nonautonomous transposition (LE *et al.* 2000; YU *et al.* 2000). Given that a low proportion of TE insertions with coding capacity are contributing to distribution patterns, the action of selection against TE-derived gene products seems unlikely to provide a primary explanation for our results.

Secondly, an important assumption of our analyses is that current estimates of recombination rates in *A. thaliana* are reflective of those in which TE accumulation occurred. If recombination rates have recently changed substantially, our ability to detect an effect of recombination on genome evolution may be compromised. While this possibility cannot be completely excluded, recent

comparative mapping in the outcrossing *Arabidopsis lyrata* (O. Savolainen, pers. comm.), which diverged from *A. thaliana* approximately 5 million years ago (KOCH *et al.* 2000), suggests that map distances are very similar in these two species, indicating that recombination rates have remained relatively stable over this time period.

Thirdly, studies of recombination hotspots in yeast and plants have provided some evidence that recombination preferentially initiates near genes (reviewed in (LICHTEN and GOLDMAN 1995; SCHNABLE *et al.* 1998). If ectopic exchange events are thus more likely to occur close to coding regions, TEs may show a pattern of under-representation in regions of high gene density under the ectopic recombination model. Because our recombination rate estimates are estimated over a large scale, we might not observe this local effect of recombination on TE abundance in our analysis. Furthermore, ectopic exchange events between elements in regions of high gene density may be more deleterious to the organism, allowing greater accumulation in regions with fewer genes. While more understanding of both the fitness effects of ectopic exchange and the relationship between ectopic exchange and recombination initiation are obviously needed, the lack of a negative correlation between recombination rate and TE abundance once gene density is factored out (from partial correlation analysis) makes these interpretations unlikely.

Testing whether ectopic exchange is of general importance, however, will require comparisons of these results with TE distribution patterns in related outcrossing species (WRIGHT *et al.* 2001), where ectopic exchange events are potentially more frequent. Recent evidence for local suppression of recombination in TE-rich intergenic regions in maize (FU *et al.* 2002) is consistent with models of

ectopic exchange, and it might suggest that low levels of ectopic exchange in the intergenic DNA have allowed for the rapid accumulation of retrotransposons (SANMIGUEL *et al.* 1998) in this species. Alternatively, however, this local reduction in recombination might imply that TE silencing by DNA methylation and changes in chromatin structure may themselves be a cause of recombination suppression (GENDREL *et al.* 2002). If so, ectopic recombination may be generally unimportant in driving TE distributions in organisms that suppress TE activity by methylation, since recombination is effectively suppressed within TE-rich regions. Variation among such organisms in the proportion of neutral insertion sites could then be more important than recombination in driving striking differences in the abundance of transposable elements. For example, recent polyploid evolution or high rates of gene duplication could create a large number of functionally redundant targets for TE insertion, leading to relaxed selection against insertions near duplicate genes (MATZKE *et al.* 1999). However, it is unclear whether the action of selection against deleterious insertions can by itself lead to a stable equilibrium in element abundance (CHARLESWORTH and LANGLEY 1989), and other forces may also be important in controlling TE copy number. Possible additional factors include the presence of self-regulated transposition, or stochastic loss of active elements, both of which may be more frequent in highly inbred and asexual populations (CHARLESWORTH and LANGLEY 1996; MORGAN 2001; WRIGHT and FINNEGAN 2001; WRIGHT and SCHOEN 1999). Nevertheless, the data indicate that selection against the deleterious effects of the disruption of gene expression may be a general force driving patterns of TE distribution, and they suggest the importance of analyses to uncouple recombination rate and gene density effects in other eukaryotic genomes. Finally, although our

results suggest that TE insertions near genes are generally under purifying selection, this does not rule out an important role of some TE insertions in the evolution of gene expression (DABORN *et al.* 2002), and unusual TE insertions in promoter and coding regions represent potential candidates for adaptive evolution.

Chapter 6

Relationship between recombination rate and nucleotide diversity in a natural population of *Arabidopsis lyrata*

6.1 Introduction

Recombination rates are expected to influence levels of DNA sequence diversity across the genome, due to the action of both positive and negative selection at linked sites (ANDOLFATTO 2001; BARTON 2000; BRAVERMAN *et al.* 1995; CHARLESWORTH 1996; KIM and STEPHAN 2000). There is consistent evidence for a strong correlation between recombination rate and DNA diversity in *Drosophila melanogaster* (ANDOLFATTO and PRZEWORSKI 2001; BEGUN and AQUADRO 1992), suggesting the importance of natural selection in structuring genome-wide patterns of polymorphism. However, in other organisms investigated, the evidence is less clear. In humans, a positive correlation between recombination and diversity is observed (NACHMAN 2001; NACHMAN *et al.* 1998), but there is a similar relationship between recombination rate and synonymous divergence (HELLMANN *et al.* 2003; LERCHER and HURST 2002), suggesting that neutral differences in mutation rate explain the pattern, rather than the effects of natural selection. Similarly, there is a strong correlation between single nucleotide polymorphism (SNP) density in noncoding regions and recombination rate in *C. elegans*, but there is also a strong correlation with silent divergence (CUTTER and PAYSEUR 2003), making it difficult to uncouple the effects of mutation and natural selection.

In plants, the influence of recombination rate on DNA sequence diversity has been investigated in detail in both tomato (BAUDRY *et al.* 2001) and maize (TENAILLON *et al.* 2002). No correlation was observed between recombination

nodule estimates of recombination rate and nucleotide diversity in maize, although estimates of recombination rate using patterns of linkage disequilibrium were found to correlate with levels of nucleotide variability (TENAILLON *et al.* 2002). Given the lack of evidence for an effect of physical recombination rates, this latter correlation most likely reflects neutral effects of the coalescent process and demographic history on both linkage disequilibrium and genetic diversity. However, this study included only one locus from the centromeric region of strongly reduced recombination, which may have influenced the power to detect a correlation. In tomato, a consistent positive correlation was observed between recombination rate and diversity, but the relationship is weak and non-significant for several tests, suggesting that the action of linked selection may not be playing a major role in structuring levels of genetic diversity across the genome in these taxa (BAUDRY *et al.* 2001). This study, however, made use of recombination rate estimates from a related species, and it was therefore not possible to rule out local changes in rates of recombination across taxa.

Here, I investigate the relationship between recombination and DNA sequence diversity within an Icelandic population of *Arabidopsis lyrata*. I make use of recent comparative mapping data between *A. thaliana* and *A. lyrata* (Kuittinen et al, unpublished) to estimate rates of recombination at these loci. I chose the first chromosome from *A. thaliana* for this investigation since 1) comparative mapping suggested strong conservation of marker order and map distances between species for this chromosome, and 2) fine-scale maps of the region surrounding the centromere in *A. thaliana* (HAUPT *et al.* 2001) provided a precise delimitation of the

region of reduced recombination and an accurate estimate of the amount of recombination suppression in this region.

6.2 Methods

Data collection

Table 6.1 shows the genomic location and size for the 18 loci surveyed in this study. I used a direct diploid sequencing strategy, using single large exons for PCR amplification and sequencing to avoid insertion/deletion (indel) polymorphisms. Coding sequences were chosen from four genomic regions of contrasting recombination rate along chromosome 1 (see Figure 6.1 and below). Exons from across chromosome 1 of *A. thaliana* were selected for direct sequencing in *A. lyrata*, using the annotation from the *A. thaliana* genome project. Exons were chosen to be greater than 800bp, and to encode genes with at least one full-length cDNA sequenced to ensure correct annotation. Exons were submitted to a BLAST search against the 'non-redundant' (nr) database in Genbank, to confirm that they are single-copy in *A. thaliana*. Exons were also submitted to a BLAST search of the genomic survey sequence (gss) database, to check for the presence of orthologous regions in the shotgun genome sequence of *Brassica oleraceae*. These orthologous regions were aligned to the *A. thaliana* genomic sequence, to identify conserved regions for primer design. PCR primers were designed with the aid of PrimerQuest (Integrated DNA technologies, http://biotools.idtdna.com/primerquest/), to amplify 650-750 base pair fragments for sequencing using the same forward and reverse primers. Sequencing reactions were conducted using the BigDye sequencing kit, and sequences were run on an ABI 3700 capillary sequencer. Chromatograms were

analysed and managed using Sequencher 4.0.5, using the 'call secondary peaks' option to aid in the identification of heterozygous sites. All chromatograms were

Table 6.1- loci surveyed in this study

locus	size of aligned region (bp)	Position along chromosome 1 (bp)	region	recombination rate A. thaliana *(cM/MB)*	recombination rate, A. lyrata *(cM/MB)*
AT1G01040	549	23519	1	0	0.413
AT1G06520	563	1995056	1	2.7	1.5
AT1G06530	552	2001642	1	2.7	1.5
AT1G10900	532	3632696	1	3.9	2.5
AT1G10980	546	3667717	1	3.9	2.6
AT1G11050	544	3681913	1	4.0	2.6
AT1G15240	576	5301823	1	4.8	3.3
AT1G19800	605	6846827	1	5.2	3.8
AT1G23200	566	8227250	1	5.3	4
AT1G36310	589	13682846	2	0.1	0.07
AT1G36370	556	13710674	2	0.1	0.07
AT1G36730	590	13914227	2	0.1	0.07
AT1G62310	638	22744866	3	4.8	11.9
AT1G62390	598	22795574	3	4.8	11.8
AT1G62520	584	22853193	3	4.8	11.7
AT1G64170	695	23525136	3	4.8	10.3
AT1G68520	603	25418251	4	4.8	6.8
AT1G68530	572	25422016	4	4.8	6.7

carefully checked manually for heterozygous nucleotide positions, using the sequence from both strands to confirm putative heterozygous sites.

Recombination rate estimation

Genetic and physical distances of markers from chromosome 1 of *Arabidopsis thaliana*, and polynomial-based estimates of recombination rates, were generated as previously described (Chapter 4). In addition, I used the genetic map of *Arabidopsis lyrata* (Kuittinen et al, unpublished) to assess the similarity in recombination rates between species. Using the *A. thaliana* genome project sequence, the physical positions of all *A. lyrata* genetic markers homologous to regions of *A. thaliana* chromosome 1 were extracted, and the relationship between

physical distance in *A. thaliana* and genetic distance in *A. lyrata* was examined. By plotting this relationship, the similarity in both gene order and recombination rates can be assessed. Estimates of recombination rate in *A. lyrata* can then be extracted in a similar way to those in *A. thaliana*, by fitting third-order polynomials to the relationship between physical and genetic distance, and taking the first derivative of this polynomial at a given physical position. Note that this analysis relies on the assumption that physical distances are similar between the two species. Several genome size estimates have been made in *Arabidopsis lyrata*, and the consensus view is that the genome is approximately 1.5 times larger than *A. thaliana* (T. Mitchell-Olds, B. Mable, O. Savolainen, pers. comm.). If this difference in size is distributed uniformly across the genome, this will not influence estimates of relative rates of recombination across the genome, and absolute rates can be calculated simply by dividing by the 1.5-fold difference in genome size. While this latter assumption needs to be tested further, the analysis in Chapter 2 showed a consistent increase in intron length in *Arabidopsis lyrata* compared with *A. thaliana*, suggesting that the difference in genome size may be due to a general difference in the relative rates of insertion and deletion, which would be equally distributed across the chromosome. I therefore simply divide all polynomial estimates by 1.5 to get estimates of recombination rate for each locus and genomic region on the chromosome arms in *Arabidopsis lyrata* (shown in Table 6.1). For the region of reduced recombination surrounding the centromere (region 2 in Figure 6.1), marker density is low in the mapping data for both species. I therefore used recombination rate estimates from the fine-scale centromere mapping of chromosome 1 of *A. thaliana* (HAUPT *et al.* 2001), and assumed an equivalent relative suppression of

recombination rates in *A. lyrata* (see below). To investigate the effect of recombination rate on levels of DNA diversity, I take two main approaches; 1) comparison of diversity levels across large-scale genomic regions of contrasting recombination (see below), and 2) correlations between diversity at each locus and point estimates of recombination rate using the polynomial fits along the chromosome arms and the fine-scale mapping near the centromere.

Data analysis

Average numbers of pairwise differences within *A. lyrata* (π), and pairwise estimates of the number of synonymous (K_s) and nonsynonymous (K_a) substitutions per site were calculated using DnaSP. version 3.95. To compare diversity levels across regions with contrasting recombination rates, I used a maximum likelihood estimate of $\theta=4Neu$ using the number of segregating sites (Chapter 4), using the approximate 95% credibility intervals assuming the chi square distribution.

Maximum likelihood test of the standard neutral model

The HKA test evaluates the fit of multilocus polymorphism and divergence data to the standard neutral model, and we use the multilocus HKA program of J. Hey to test this. However, while the standard multilocus HKA test gives a general test for the fit of the data to a neutral model, it does not allow for an explicit test of selection on a particular locus. Furthermore, the HKA test does not provide a method to explicitly compare different models, for example in order to test for significant heterogeneity in mutation rates across loci. I therefore used a maximum likelihood analysis of synonymous polymorphism and divergence, derived from the

HKA framework, making use of the polymorphism and divergence data. Similarly to the HKA test, the method assumes no recombination within loci, free recombination (independence) between loci, and it assumes that the ancestral population size is the same as the current population size. Under neutrality, the likelihood of the observed numbers of segregating sites (S_i) and pairwise divergence (D_i) across each locus i, for a total of r loci, is given by:

$$(1)\quad L = \prod_{i=1}^{r} L_i(\theta_i / S_i) L_i(\theta_i, T / D_i)$$

Where $\theta_i=4Nu_i$, N is the effective population size, u_i is the locus-specific mutation rate, and T is the divergence time of the two species, in units of $2N$ generations. Also similarly to HKA, equation (1) assumes that polymorphism and divergence are independent. Although a slight positive correlation is between S_i and D_i is expected, the effect is expected to be very weak, and violations from the independence assumption lead to slightly conservative tests (HUDSON *et al.*, 1987). Under the assumption of no recombination, the likelihood of θ_i given S_i is given by (HUDSON 1990 Chapter 4):

$$(2)\quad L_n(\theta_i / S) \propto P_n(S / \theta_i) = \sum_{i=0}^{S} P_{n-1}(S - i / \theta_i) Q_n(i / \theta_i)$$

where

$$P_2(S / \theta) = \left(\frac{\theta}{1+\theta}\right)^S \frac{1}{1+\theta}$$

and

$$Q_n(S/\theta) = (\frac{\theta}{\theta+n-1})^S \frac{n-1}{\theta+n-1}$$

The likelihood of the number of pairwise differences between a random sequence chosen from each species is given by (TAKAHATA *et al.* 1995):

$$(3)\ L(T,\theta_i/D) \propto P(D/\theta_i,T) = (1-\frac{\theta_i}{1+\theta_i})(\frac{\theta}{1+\theta})^D e^{-\theta_i T} \sum_{m=0}^{D} (\frac{(1+\theta_i)\theta_i T}{\theta_i})^m$$

The full neutral model thus has r+1 parameters; a θ parameter for each locus, plus the shared divergence time parameter. Results using coalescent simulations suggest that this likelihood method generates good parameter estimates under the standard neutral model (See Appendix A.3). To test for heterogeneity in mutation rates, I compared a fixed-rate model, where $\theta_1=\theta_2=..=\theta_n$, with a free-rate model which allows each locus its own value of θ. Significance was assessed using a likelihood ratio test, using the chi-square distribution with r-1 degrees of freedom.

I also used the likelihood ratio to test for the action of natural selection at a candidate gene, here defined as the rth locus. In the HKA framework, selection acts by uncoupling polymorphism from divergence; positive selection will reduce estimates of θ from polymorphism data relative to divergence data, while balancing selection will cause an excess of diversity relative to divergence. The selection model thus has one additional parameter $k=\frac{\hat{\theta}}{\theta_0}$, where $\hat{\theta}$ is the maximum likelihood population mutation parameter estimated using S_r alone, and θ_0 is the expected value of $4N_e u$ at this locus under the multilocus neutral model:

$$(4)\ L = L(k\theta_r / S_r)L(\theta_r, T / D_r)\prod_{i=1}^{r-1} L(\theta_i / S_i)L\ (\theta_i, T / D_i)$$

In this model, k measures the degree to which diversity is increased or decreased by the action of selection at locus r. By calculating the maximum likelihood under both the selection and the neutral model, selection at locus r can be tested using a likelihood ratio test, where twice the difference in log-likelihood is approximately chi-squared distributed with one degree of freedom. Simulation results suggest that the test remains approximately chi squared distributed assuming realistic levels of intragenic recombination (see Appendix A.3). To maximize the likelihood of these two models, I use a Monte-Carlo Markov chain using the Metropolis-Hastings algorithm. Briefly, a parameter is chosen at random, and the value of this parameter is incremented using a uniform distribution between $-\lambda$ and $+\lambda$, where λ is a predefined increment. The likelihood of the data under the new parameter combination is then calculated, and if the likelihood of the new parameter set is greater, the change is accepted. If the likelihood is less than the previous likelihood, the change is accepted with probability proportional to the difference in log likelihoods:

(5) $P[accept]=L_{new}/L_{old}$

Because of the large number of parameters and our primary interest in the maximum likelihood, I follow a similar approach of simulated tempering to (McVean and Vieira 2001); for changes with lower likelihood, the probability of acceptance is determined by multiplying the difference in log likelihood by a factor of 50. After

10,000 iterations, this acceptance probability is reduced to a factor of 0.5 times the difference in log likelihoods for 1,000 iterations, and then returned to the original value. The Markov chain is run multiple times using different starting parameters and different random number seeds to ensure convergence.

6.3 Results and Discussion

6.3.1 Comparison of recombination rates in A. thaliana first chromosome vs. A. lyrata

Figure 6.1 shows the relationship between physical distance along the *A. thaliana* chromosome 1 and genetic distances in *A. thaliana* and *A. lyrata*. Markers orthologous to the *A. thaliana* chromosome 1 have been shown by Kuittinen *et al.* (unpublished) to fall into two linkage groups in *A. lyrata*, suggesting the presence of a chromosome fusion in *A. thaliana*. Chromosome painting has suggested that this *A. lyrata* karyotype is the same in *C. rubella* (M. Lysack, *pers. comm.*), providing confirmation that there is a derived chromosome fusion in *A. thaliana*. With the exception of this putative fusion event, there is little evidence to suggest the presence of extensive rearrangements between the two species; marker order is highly conserved between the two species, and regional differences in genetic map distances appear to be conserved (Figure 6.1). In particular, despite low marker density, there is evidence for a decline in the rate of recombination in the region corresponding to the *A. thaliana* centromeric region (region 2), suggesting that the location of the centromere and associated recombination suppression are conserved in this region. Recent data from chromosome painting in *Arabidopsis lyrata* has provided further evidence that the position of the centromere is conserved in this

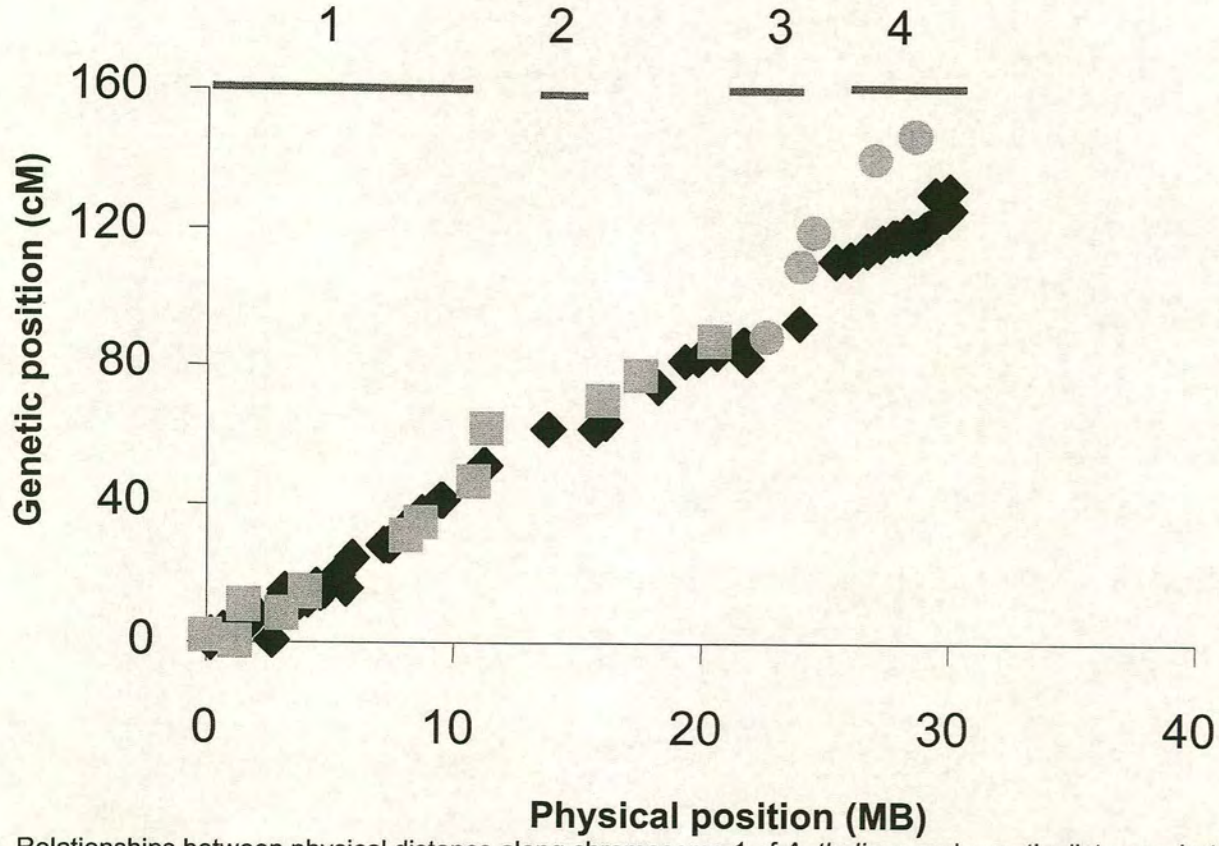


Figure 6.1. Relationships between physical distance along chromosome 1 of *A. thaliana* and genetic distances between markers in *A. thaliana* (black) and *A. lyrata* (grey). Squares and circles represent the two linkage groups in *A. lyrata*. Physical and genetic distances between the two *A. lyrata* linkage groups do not have meaning, but they are aligned continuously for comparison with *A. thaliana*. Numbers above represent the four genomic regions surveyed for polymorphism in *A. lyrata*.

region (M. Lysack, *pers. comm.*). One exception to the general correspondence in rates of recombination is evidence for a higher rate of crossing over in *A. lyrata* in the second linkage group (region 3 in Figure 6.1). This is most likely due to the fact that this is a shorter chromosome arm in *A. lyrata*, which tends to generate a higher rate of crossing-over per unit of physical distance (KABACK *et al.* 1999). Table 6.1 shows polynomial based estimates of recombination rates along the chromosome arms for each locus, and the estimates of recombination rate in the centromeric regions from fine-scale mapping in *A. thaliana*. The estimates of recombination rate along the chromosome arms in *A. lyrata* show a moderately strong and significant positive correlation with estimates from *A. thaliana* (Spearman's rank correlation=0.63, $p<0.05$, n=15), confirming the similarity in map distances between the two species.

6.3.2 Effects of recombination rate on neutral genetic diversity

Table 6.2 gives a summary of levels of synonymous and replacement nucleotide diversity and divergence across the 18 loci surveyed in this study. There is considerable scatter across individual loci in levels of diversity; this is at least partially due to the small number of synonymous sites in the coding fragments sequenced. Nevertheless, with the number of loci sequenced, joint information using multiple loci provides extensive information on levels of diversity in different genomic regions. Across all loci, the maximum likelihood estimate of synonymous θ conditioning on the number of segregating sites is 0.008. This estimate is lower than previous estimates of within-population silent θ in Scotland and another population from Iceland, but is still considerably higher than joint maximum

likelihood estimates of θ within populations from North America (Chapter 4). Subdividing the data into large genomic regions, there is evidence for significant differences across regions in levels of neutral genetic diversity (Figure 6.2). However, in contrast with predictions, there is no evidence for a reduction in variability in the pericentromeric regions with highly suppressed recombination rates; in fact, variability is significantly higher in the centromeric region than in two out of three genomic regions on the chromosome arms (Figure 6.2). Using point estimates of recombination rate for each locus, there is no significant correlation between

Table 6.2- Summary of nucleotide polymorphism across loci

locus	n	L_s[1]	L_n[1]	S_s	S_n	π_s	π_n	K_s	K_a
AT1G01040	18	117	411	3	1	0.011	0.0003	0.162	0.012
AT1G06520	20	122	427	0	1	0	0.0009	0.153	0.061
AT1G06530	17	108	441	1	1	0.0045	0.0011	0.141	0.049
AT1G10900	22	122	406	7	3	0.013	0.0016	0.092	0.008
AT1G10980	8	123	420	0	0	0	0	0.108	0.032
AT1G11050	17	132	402	0	2	0	0.0006	0.189	0.024
AT1G15240	20	120	439	0	0	0	0	0.117	0.028
AT1G19800	21	152	448	0	1	0	0.0011	0.148	0
AT1G23200	4	136	426	5	3	0.018	0.0035	0.147	0.039
AT1G36310	20	125	454	26	6	0.036	0.0033	0.409	0.018
AT1G36370	21	120	414	6	1	0.017	0.0004	0.178	0.015
AT1G36730	24	119	448	0	0	0	0	0.283	0.02
AT1G62310	18	136	497	1	2	0.0015	0.0008	0.168	0.032
AT1G62390	15	125	469	6	5	0.014	0.0033	0.122	0.013
AT1G62520	18	134	427	6	2	0.014	0.0015	0.247	0.028
AT1G64170	16	130	362	9	1	0.022	0.0004	0.181	0.022
AT1G68520	4	117	471	0	0	0	0	0.182	0.057
AT1G68530	18	132	438	0	0	0	0	0.119	0.010

[1], numbers of synonymous and nonsynonymous sites

recombination rate and diversity using both synonymous θ and π (p>>0.05; Figure 6.3A). This suggests that beneficial mutations and recurrent purifying selection are not causing a reduction in variability near the centromere, where recombination is highly suppressed.

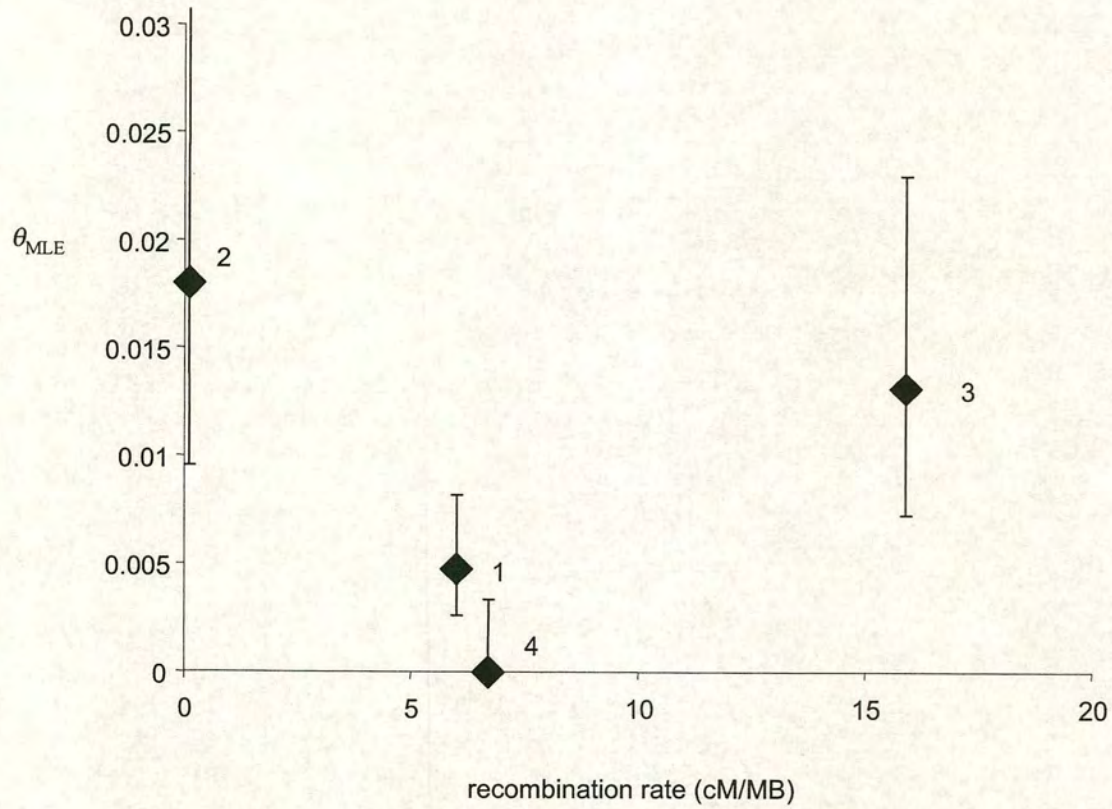


Figure 6.2. Recombination rate vs. diversity across the four genomic regions surveyed. Numbers represent the corresponding regions shown in figure 6.1. Values are joint maximum likelihood estimates of synonymous θ given the number of segregating sites across the loci surveyed, with 95% credibility intervals shown. Recombination rates for each genomic region are estimated by taking the slope of physical vs. genetic distance across the entire region.

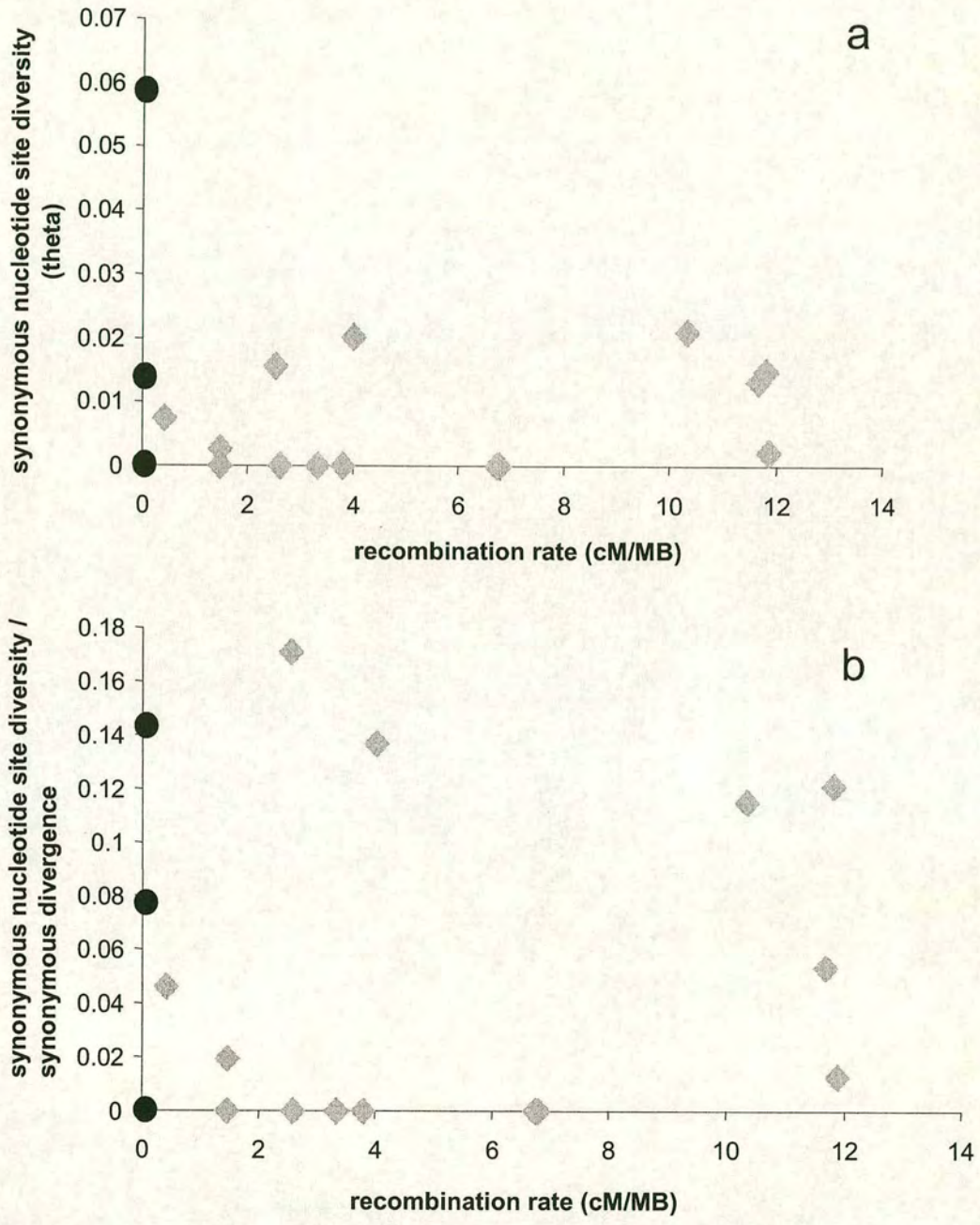


Figure 6.3. Point estimates of recombination rate vs. a) nucleotide diversity, as measured by synonymous theta, and b) synonymous diversity relative to divergence, calculated using synonymous theta divided by Ks. Loci from the centromeric region are shown in black, while loci from the chromosome arms are in grey.

6.3.3 Testing for heterogeneity in mutation rates

One possible explanation for a lack of reduction in diversity in the low-recombination region is heterogeneity in mutation rates across loci. If there are differences across loci and genomic regions in mutation rate, this will generate noise when testing for an effect of recombination on neutral variability. Indeed, the most polymorphic locus (AT1G36310) is found in the pericentromeric region 2, and shows both elevated synonymous diversity and elevated synonymous divergence (Table 6.2), suggesting that real differences in mutation rate may be contributing to differences in nucleotide diversity across the genome. This locus also shows highly elevated divergence between *A. thaliana* and *Brassica oleracea* (Ks=0.82, average Ks between *A. thaliana* and *Brassica* spp.: 0.474, TIFFIN and HAHN 2002). Across all loci synonymous θ shows a highly significant correlation with divergence ($r=0.6139$, $p<0.01$), as expected if mutation rates differ across loci. Using maximum likelihood, a model that allows for locus-specific differences in mutation rate (Model 2) provides a highly significant improvement to the likelihood compared with a model of fixed mutation rate (Model 1), providing confirmation of the hypothesis of heterogeneity in mutation rates across loci (Table 6.3). These results highlight the importance of incorporating mutation rate variation into parameter estimation (e.g. Wall, 2003). Figure 6.4 shows estimates of θ from both maximum-likelihood and conventional HKA tests of polymorphism and divergence; the pattern is suggestive of an increase in mutation rate close to the centromere. This result contrasts with recent analyses in other taxa suggesting a positive correlation between recombination and mutation rates (CUTTER and PAYSEUR 2003; HELLMANN *et al.* 2003).

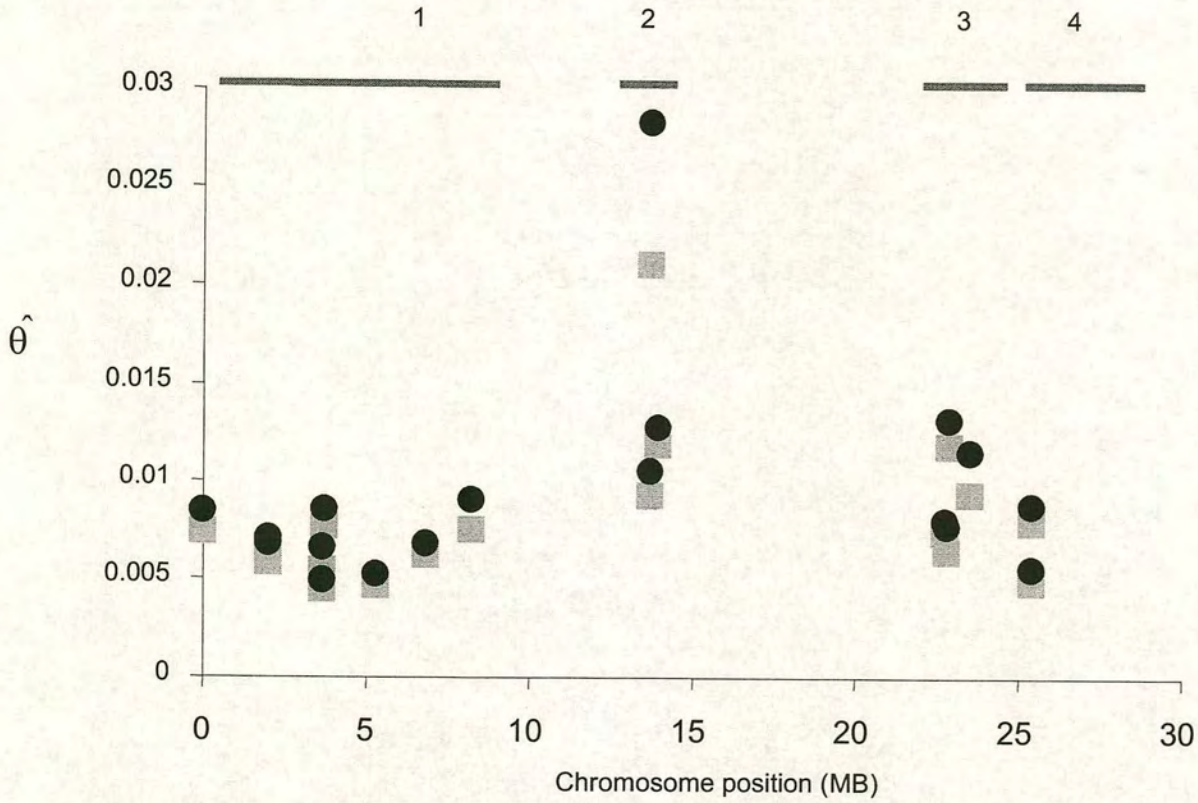


Figure 6.4- Effect of chromosome position on estimates of synonymous θ from joint polymorphism and divergence data. Black circles are parameter estimates using the standard multilocus HKA test, and grey boxes are maximum likelihood estimates under the full neutral model. Regions of contrasting recombination rate, corresponding to Figure 1, are shown above.

To correct for these differences in mutation rate, I calculated the partial correlation between synonymous diversity (θ) and point estimates of recombination rate, corrected for divergence. There is still no significant correlation between diversity and recombination rate when correcting for divergence (partial r=0.03, p>>0.05) suggesting that mutation rate heterogeneity alone cannot explain the observed pattern. Figure 6.3B shows a plot of the relationship between recombination and the ratio of diversity to divergence for each region; clearly, there is still no indication of reduced variability in the central region with highly suppressed recombination.

6.3.4 Unusual selection acting in the centromeric region

Another possibility is that the rate and strength of natural selection are not uniform across genomic regions. Models predicting a reduction in genetic diversity in regions with reduced recombination assume that selection is similar across the genome, but there are several ways in which this assumption may be violated. First, gene density and gene expression level are substantially reduced in the pericentromeric regions (Chapter 5). Given this, there may simply be too few targets for either positive or negative selection in these regions to cause a reduction in genetic diversity. On the *A. thaliana* chromosome 1, the central regions of suppressed recombination only contain approximately 4% of the coding sequence along the chromosome, and only 1% of the total coding sequence in the genome. Of this coding sequence, a lower percentage has evidence for expression than in other regions, and there is likely to be a higher fraction of pseudogenes. This contrasts with the genome structure of *Drosophila melanogaster*, where a much higher

fraction of the central regions of autosomes have suppressed recombination, and a strong correlation between recombination rate and diversity is predicted (CHARLESWORTH 1994).

Alternatively, the action of balancing selection could enhance variability near the centromere, as has been suggested in Lycopersicon (BAUDRY *et al.* 2001). In particular, despite the high mutation rate, the ratio of synonymous diversity relative to synonymous divergence is among the highest at the centromeric locus AT1G36310 (Figure 6.3A), suggesting the possibility of the action of selection maintaining variability at or near this locus. A multilocus HKA test across all 18 loci shows a highly significant departure from neutral expectation ($p<10^{-3}$), with AT1G36310 polymorphism showing the largest contribution to the chi-square (deviation=5.6). To test further for the specific action of selection on AT1G36310, we used a maximum likelihood analysis, comparing a neutral model with a model allowing for selection to act at this locus (Table 6.3). Consistent with the hypothesis of selection, the latter model provides a significant improvement to the likelihood for both the fixed (Model 3) and free (Model 4) mutation-rate models. Furthermore, a model allowing both selection at this locus and free mutation rates across loci shows a significant improvement to the likelihood compared with both the selection model with fixed mutation rates and the neutral model with free mutation rates, providing support for both heterogeneity in mutation rates and the action of natural selection elevating variability. Under the free mutation rate model, the maximum likelihood estimate of the selection parameter k is 4.3, suggesting a fourfold increase in variability relative to expectation.

Given that our analysis suggests that both selection and mutation processes may be unusual in the pericentromeric regions, loci from these regions can be removed from the analysis, to investigate the possibility of a correlation between recombination rate and diversity when the centromeric regions are excluded. Excluding region 2, the relationship between recombination rate and θ is positive, but not significant (Pearson r=0.36, p>>0.05). This suggests that unusual effects of selection in the region surrounding the centromere cannot alone explain the lack of correlation between recombination rate and diversity. Although increasing the number of loci may increase the power to detect an effect along the chromosome

Table 6.3- maximum likelihood analysis of synonymous polymorphism and divergence

model	description	k^2	L	model comparison	likelihood ratio statistic (d.f.)
1:	fixed mutation no selection	-	-131.8		-
2:	free mutation no selection	-	-95.9	M2 vs. M1	71.8*** (17)
3:	fixed mutation selection	9.1	-119.8	M3 vs. M1	24*** (1)
4	free mutation selection	4.3	-91.4	M4 vs. M3 M4 vs. M2	56.8***(17) 9** (1)

1, Selection parameter for the gene AT1G36310
**, p<0.01; ***, p<0.001

arms, our results suggest that recombination rate does not have a strong effect on diversity in this population.

6.3.5 Influence of demographic history

Recent changes in population size and population dynamics could also contribute to an uncoupling of recombination from diversity, and obscure the signature of selection. Previous analysis has shown elevated linkage disequilibrium in Northern European populations of *Arabidopsis lyrata* (Chapter 4), suggesting the possibility of a recent population bottleneck, or population admixture associated with postglacial colonization. Furthermore, analysis of species-wide variability in *A. lyrata* and *A. halleri* suggests the possibility of ongoing gene flow and/or recent introgression between these species (Ramos Onsins *et al.*, submitted). If there has been population admixture or introgression between species, this could inflate variability in regions of reduced recombination relative to equilibrium conditions, due to differentiation at these regions between ancestral populations. Thus, recent demographic history could mask the signature of historical selection, making it difficult to test models of genetic hitchhiking. Although our maximum likelihood analysis showed evidence for an effect of both natural selection and heterogeneity in mutation rates, the analysis was conducted assuming the standard neutral model, without incorporating subdivision or demographic history. Both population subdivision and recent departures from demographic equilibrium can influence the ratio of diversity to divergence (WAKELEY 2003). Although the effects of these processes on tests of neutrality have not fully been explored, the distribution of estimates of divergence and diversity are also likely to be affected.

In our previous analysis of linkage disequilibrium (Chapter 4), it was difficult to rule out the possibility of lower recombination rates in *A. lyrata* compared with *A. thaliana* as an alternative explanation to demographic history for

the observed excess LD. With the availability of comparative mapping data from *A. lyrata*, this hypothesis can be tested explicitly. For syntenic markers between the two species, the *A. lyrata* genetic map shows a total length of 258.7 cM, and a total physical distance in *A. thaliana* of 122.6 Mb. Assuming that the genome size of *A. lyrata* is 1.5 times larger, this would imply that the average rate of crossing over is 2.8 cM/Mb in *A. lyrata*, or slightly more than half the genome average of 5 cM/Mb in *A. thaliana*. Thus, if we assume a uniform elevation of genome size across the *A. lyrata* genome, there is evidence for an elevation of per base pair rates of crossing over in *A. thaliana* compared with *A. lyrata*, as expected under some models for the evolution of recombination rates in self-fertilising species (CHARLESWORTH and CHARLESWORTH 1979). However, this estimated difference is unlikely to explain the observed excess linkage disequilibrium in *A. lyrata*; using our five loci we estimated an average of a seven-fold reduction in the effective rate of recombination (Chapter 4) compared with expectation.

While the above results suggest that a global reduction in rates of recombination are unlikely to explain the excess LD, it is possible that the loci surveyed in Chapter 4 reside in regions of locally reduced recombination in *A. lyrata*. However, we find no evidence from the comparative mapping data that this is the case; three of the five nuclear loci are located in regions of synteny between the two species, and point estimates of recombination rate for two of the loci are close to the genome-wide average (CAUL, 2.9 cM/Mb; HAT4, 2.8 cM/Mb), and for a third it is substantially higher than average (8.6 cM/Mb). Thus, we find no evidence for unusual suppression of recombination at these loci, and our original suggestion of an excess of linkage disequilibrium remains when estimates of recombination rate are

made using the comparative mapping data. Although marker density remains fairly low in the comparative map, making it possible that we are missing more localized suppression of recombination, these results suggest that demographic history is a much more likely explanation than the evolution of recombination rates for the observed patterns of LD.

Given strong evidence for an important departure from demographic equilibrium, this effect is likely to play an important role in uncoupling recombination from diversity in these populations. To test this model further, it will be important to investigate the relationship between recombination and diversity in central European populations of *A. lyrata*, which may have experienced less demographic perturbation, and may therefore be closer to equilibrium. Combined with the previous analysis of variability and linkage disequilibrium in *A. lyrata*, the results suggest that demographic history is playing a more important role than natural selection in structuring genomic patterns of genetic variability in the populations studied, although further analysis of variability surrounding the centromere will be important to test for the presence of a balanced polymorphism in this chromosomal region.

Chapter 7 Conclusions

7.1 Efficacy of natural selection in selfing vs. outcrossing Arabidopsis

A central goal of this thesis was to test for a reduction in the efficacy of selection in the self-fertilising *Arabidopsis thaliana*, in comparison with its closely related, outcrossing congener *A. lyrata*. Comparisons of gene structure and rates of substitution revealed two consistent differences between species; intron size is significantly larger in *A. lyrata* (Chapter 2), and the GC content at synonymous sites is lower in *A. thaliana* (Chapter 3). The first result contrasts with a priori expectations that longer introns are slightly deleterious and should accumulate in a self-fertilising species, and suggests the possibility of directional selection on intron length in a weedy annual. While the second difference is consistent with a reduction in levels of codon usage bias in *A. thaliana*, more direct comparisons of the frequency of preferred codons based on tRNA gene abundance showed no consistent differences between species (Chapter 3), and the relative rates of preferred and unpreferred codons were similar between the species (Chapters 2 and 3). Assuming that tRNA gene abundances are conserved across species, this result would suggest a neutral explanation for the difference in GC content, rather than differences in the efficacy of natural selection. Similarly, the relative rates of amino acid substitution are very similar between the two species (Chapter 3) and, using this set of loci, the levels of amino acid polymorphism relative to synonymous polymorphism are also similar. Thus, in contrast with previous suggestions (BUSTAMANTE *et al.* 2002), there is no evidence from these results for a strong accumulation of deleterious

mutations in *A. thaliana* as a result of high levels of inbreeding, and we find no molecular genetic evidence for a rapid decline in fitness in this selfing species.

7.2 Testing models of inbreeding and the effects of history

One obvious limitation of this study is that it is only a comparison of one outcrossing species with one self-fertilising species, and so in a sense represents a single data point. Ideally, one would like to compare a large number of independent pairs of species with contrasting breeding systems. In addition to the obvious technical challenge of collecting large amounts of genomic data for many species, there is also the problem of identifying suitable species pairs; in many cases selfing species are very closely related to their outcrossing congeners, giving little time for detectable changes in molecular evolution.

Nevertheless, because genes can evolve independently, and the presence of recombination means that the study of genome-wide polymorphism data gives information on independent resolutions of coalescent history, a genome-wide analysis of two taxa does allow for powerful inferences to be made. The ability to test models of effects of inbreeding from such a comparison, however, is dependent on how well the species conform to population genetic models, i.e. whether they share the same population genetic parameters and are at equilibrium. The study of nucleotide polymorphism in four populations of *A. lyrata* (Chapter 4) provided evidence for important departures from these assumptions in this species, including a recent population bottleneck in North America, and an excess of linkage disequilibrium in European populations. Subsequent analysis making use of comparative mapping data suggested that this result is most likely the result of

demographic history, rather than the evolution of recombination rates (Chapter 6). Despite evidence for a higher long-term effective population size in *A. lyrata*, these results suggest that the details of demographic history in both *A. lyrata* and *A. thaliana* are very important in structuring patterns of diversity in these species, and may in turn have influenced the effectiveness of selection on slightly deleterious mutations. Similarly, recent demographic history may have led to an uncoupling of the predicted relationship between recombination rate and nucleotide diversity in *A. lyrata* (Chapter 6), although a balanced polymorphism in a centromeric region may also explain the data. In general, the results suggest that recent demographic history may be more important than selection in structuring patterns of genetic diversity in these populations, and the mating system may be relatively unimportant in determining patterns of molecular evolution. Preliminary evidence for equilibrium patterns of selection on codon usage bias in species-wide samples in *A. lyrata*, but not single population samples (Chapter 3), is consistent with a role of subdivision in influencing selection at the genome level.

Based on these results, it would be useful to identify populations that have been less subject to recent departures from demographic equilibrium. In *A. lyrata*, it is possible that central European populations represent older, pre-glacial populations, and may therefore be more suitable for testing models of selection. It will therefore be important to extend genome-wide surveys of nucleotide variability to more populations of *A. lyrata*, to test the hypothesis that the effective recombination rates are closer to expected in central Europe. On the other hand, the results imply that demographic history may generally be more important than normally assumed in models of molecular evolution, and it is possible that few

natural populations actually conform to models considering unstructured panmictic populations at long-term demographic equilibrium. This means that simple models of natural selection may not be generally applicable, perhaps particularly in plant populations, where strong population subdivision and postglacial colonization may be common. It will therefore also be important to better understand the action of selection in non-equilibrium and structured populations (e.g. WAKELEY 2003), and to model the effects of selection using realistic models of species' demographic history. Such models will require high-resolution information on patterns of nucleotide polymorphism across species' ranges, to infer the relevant demographic parameters.

7.3 Understanding selection at the genome level

Another goal of this thesis was to better understand the strength and nature of natural selection on plant genomes. Analysis of the effects of gene expression confirmed and extended previous results showing higher levels of codon usage bias in highly expressed genes (Chapters 2 and 3), and I used data on tRNA gene abundance to test models of translational selection, and to identify codon preferences which match more closely with those favoured by selection (Chapter 3). Molecular evolutionary analysis also revealed an important effect of gene expression level on rates of protein evolution (Chapter 2), and subsequent analysis showed that genes expressed more broadly are under stronger selective constraints (Chapter 3). Whether this pattern reflects true differences in constraints between proteins, or higher rates of positive selection in tissue-specific genes will require more investigation, for example by sequencing genes in a more extensive set of related

species. Finally, analysis of the complete *Arabidopsis thaliana* genome (Chapter 5) suggested an important role for the action of natural selection against transposable element insertions disrupting gene function in both introns and intergenic DNA, and the results provided little evidence for selection against ectopic recombination, as expected in a highly self-fertilising organism. These patterns are consistent with previous work showing evidence for less effective selection against TEs in *A. thaliana* compared with *A. lyrata* (WRIGHT *et al.* 2001). Combined with a lack of evidence for differences in the efficacy of selection on codon bias and amino acid mutations (Chapters 2 and 3), these results suggest that there has been a relaxation in the *strength* of natural selection against TEs in *A. thaliana*, as expected if selection against ectopic exchange in playing an important role in controlling TE abundance (see Chapter 5). To test this further, it will be interesting to examine TE distributions in related outcrossers as more genomic data become available, to better understand the influence of recombination rate and mating system on the action of selection against TEs.

Appendix A.1 Genes surveyed for analysis of molecular evolution and polymorphism

locus	Description	Species	Source and Accession Number	Source for polymorphism data
At1g01030	putative DNA-binding protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
		Capsella grandiflora	this study	
At1g01040	DEAD/DEAH box helicase carpel factory	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
		Capsella grandiflora	this study	
At1g03560	pentatricopeptide (PPR) repeat-containing protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	
		Capsella grandiflora	this study	
At1g05010	1-aminocyclopropane-1-carboxylate oxidase/EFE	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	
		Capsella rubella	this study	
		Brassica oleracea	AF252853	
At1g06520	phospholipid/glycerol acyltransferase family protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
		Brassica oleracea	BH691258	
At1g06530	expressed protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
At1g10900	phosphatidylinositol-4-phosphate 5-kinase isolog	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
		Brassica oleraceae	BH921848, BZ024187	
At1g10980	membrane protein PTM1 precursor isolog	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
		Capsella grandiflora	this study	

At1g11050	protein kinase family	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
At1g12190	F-box protein family	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	AY062426	
At1g12210	disease resistance protein (CC-NBS-LRR class)	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	AY062427	
At1g12230	unknown expressed protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	AY062427	
		Brassica oleracea	BH976544, BH602574, BH957859	
At1g12240	glycosyl hydrolase family 32	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	AY062428	
		Brassica oleracea	AF274299	
At1g15240	hypothetical protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
		Capsella grandiflora	this study	
At1g19800	unknown expressed protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
		Brassica oleracea	BH647047	
At1g23200	pectinesterase family	*Arabidopsis thaliana*	NC003070	
		Arabidopsis lyrata	this study	Chapter 6
At1g25280	F-box containing tubby family protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	AY140470	
		Brassica oleracea	BH734295,BH544681,BH424212	
At1g26310	CAULIFLOWER	*Arabidopsis thaliana*	NC_003070	(PURUGGANAN and SUDDITH 1998), Zheng and Nordborg Chapter 4
		Arabidopsis lyrata	AF143381	Chapter 4
		Brassica rapa	AJ251300	
At1g36310	expressed protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6

At1g36370	hydroxymethyltransferase -related	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
		Capsella grandiflora	this study	
At1g36730	Eukaryotic translation initiation factor 5 -related	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
At1g44120	C2 domain/armadillo repeat-containing protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	
At1g55350	n-calpain-1 large subunit -related	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	
		Capsella grandiflora	this study	
At1g59720	pentatricopeptide (PPR) repeat-containing protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	
		Capsella grandiflora	thsi study	
At1g62310	expressed protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
		Brassica oleracea	BH582293	
At1g62390	octicosapeptide/Phox/Bem1p (PB1) domain-/tetratricopeptide repeat (TPR)-containing protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
		Capsella grandiflora	this study	
At1g62520	expressed protein	*Arabidopsis thaliana*	NC_003070	
		Arabdiopsis lyrata	this study	Chapter 6
		Capsella grandiflora	this study	
At1g64170	putative CHX16	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
At1g64970	gamma-tocopherol methyltransferase	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	This study	
		Brassica oleracea	AF381248	

At1g65330	MADS-box protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	AF528676	
At1g65340	putative cytochrome P450	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	AF528677	
		Brassica oleracea	BH678035,BH700230	
At1g65360	MADS-box protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	AF528679	
At1g65380	CLAVATA2/CLV2	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	AF528680	
		Brassica oleracea	BH952257	
At1g65390	putative disease resistance protein (TIR class)	*Arabidopsis thaliana*	NC_003070	
At1g65410	ABC transporter family protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	AF528683	
		Brassica oleracea	BZ514690	
At1g65420	putative antigen receptor	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	AF528684	
		Brassica oleracea	BH453439	
At1g65430	expressed protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	AF528685	
		Brassica oleracea	BH678193	
At1g65450	anthranilate N-hydroxycinnamoyl/benzoyltransferase - related	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	
At1g66340	ethylene-response protein/ETR1	*Arabidopsis thaliana*	NC_003070	Chapter 4
		Arabidopsis lyrata	Chapter 2/AF494374	Zheng and Nordborg
At1g67140	expressed protein	*Arabidopsis thaliana*	NC_003070	Chapter 4
		Arabidopsis lyrata	AY140501	

At1g68520	CONSTANS B-box zinc finger family protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
		Brassica oleracea	BZ003641	
At1g68530	very-long-chain fatty acid condensing enzyme/ CUT1	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
		Capsella grandiflora	this study	
At1g68620	expressed protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	Chapter 6
At1g69120	APETALA1/AP1	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	AF143379	
		Brassica oleracea	Z37968	
At1g72390	expressed protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	
		Capsella grandiflora	this study	
At1g74600	pentatricopeptide (PPR) repeat-containing protein	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	
		Capsella grandiflora	this study	
At1g77120	alcohol dehydrogenase/ADH	*Arabidopsis thaliana*	NC_003070	(INNAN et al. 1996)
		Arabidopsis lyrata	AF110453	
		Capsella rubella	AF110435	
At1g78850	curculin-like (mannose-binding) lectin family	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	this study	
		Capsella grandiflora	this study	
At2g04410	unknown protein	*Arabidopsis thaliana*	NC_003071	
		Arabidopsis lyrata	AY140531	
		Brassica oleracea	BH553817	
At2g16910	bHLH	*Arabidopsis thaliana*	AC005167	
		Arabidopsis lyrata	This study	
		Capsella bursa-pastoris	This study	
		Brassica napus	AJ511991	

At2g31390	putative fructokinase/FKA1	*Arabidopsis thaliana*	NC_003070	
		Arabidopsis lyrata	Chapter 2/AF494375	
		Brassica oleracea	BH55898, BH497074	
At3g07040	putative disease resistance protein/RPM1	*Arabidopsis thaliana*	NC_003074	
		Arabidopsis lyrata	AF122982	
		Brassica napus	AF105140	
At3g12500[1]	Basic endochitinase (CHIB)	*Arabidopsis thaliana*	NC_003074	(KAWABE and MIYASHITA 1999)
		Arabis gemmifera	AB023464	
At3g22060	putative receptor kinase common family	*Arabidopsis thaliana*	NC_003074	
		Arabidopsis lyrata	AY140444	
		Brassica oleracea	BZ490722, BH979133	
At3g27920	glabrous 1/GL1	*Arabidopsis thaliana*	NC_003074	
		Arabidopsis lyrata	AF263720	
		Brassica oleracea	BH595973, BZ512071	
At3g48560	acetolactate synthase	*Arabidopsis thaliana*	AL133315	
		Arabidopsis lyrata	This study	
		Capsella bursa-pastoris	This study	
		Brassica napus	M60068	
At3g51240	flavanone 3-hydroxylase/F3H	*Arabidopsis thaliana*	NC_003074	(AGUADE 2001)
		Arabidopsis lyrata	AJ295607	
		Brassica oleracea	BH568916, BH246598, BH944566	
At3g54340	APETALA3/AP3	*Arabidopsis thaliana*	NC_003074	(PURUGGANAN and SUDDITH 1999)
		Arabidopsis lyrata	AF143380	
		Brassica oleracea	U67456	
At3g55120	chalcone isomerase/CHI	*Arabidopsis thaliana*	NC_003074	(KUITTINEN and AGUADE 2000)
		Arabidopsis lyrata	AJ287322	
		Brassica oleracea	BH449595	
At4g01600	putative ABA-repsonsive protein	*Arabidopsis thaliana*	NC_003075	
		Arabidopsis lyrata	Chapter 2/AF494370	

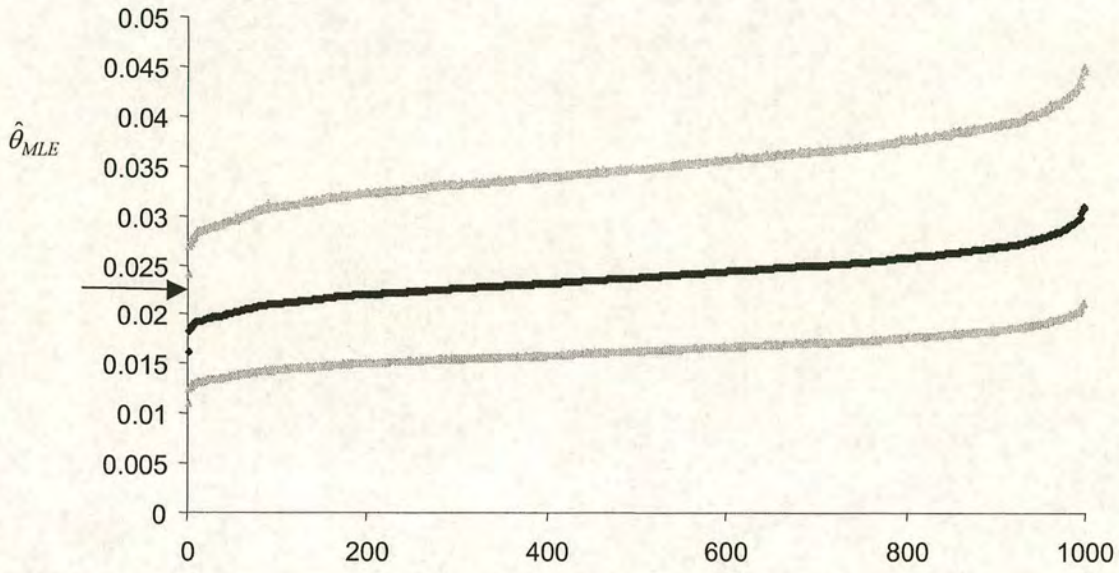
At4g01660	putative ABC transporter	*Arabidopsis thaliana*	NC_003075	
		Arabidopsis lyrata	Chapter 2/AF494372	
At4g02780	copalyl diphosphate synthase (CPS) /GA1	*Arabidopsis thaliana*	NC_003075	
		Arabidopsis lyrata	H. Kuittinen & O. Savolainen, unpublished	
		Brassica oleracea	H. Kuittinen & O. Savolainen, unpublished	
At4g03050	2-oxoglutarate-dependent dioxygenase 3/AOP3	*Arabidopsis thaliana*	NC_003075	
		Arabidopsis lyrata	AF418280	
At4g05530	Short-chain alcohol dehydrogenase/Sc-ADH	*Arabidopsis thaliana*	NC_003075	Chapter 4 Zheng and Nordborg
		Arabidopsis lyrata	Chapter 2/AY174527	Chapter 4
		Brassica oleracea	BOIFP57TR***	
At4g10180	Deetiolated1 light signal transduction protein/DET1	*Arabidopsis thaliana*	NC_003075	
		Arabidopsis lyrata	H. Kuittinen & O. Savolainen, unpublished	
		Brassica oleracea	H. Kuittinen & O. Savolainen, unpublished	
At4g16280	Flowering time control protein/ FCA	*Arabidopsis thaliana*	NC_003075	
		Arabidopsis lyrata	H. Kuittinen, unpublished	
		Brassica oleracea	H. kuittinen, unpublished	
At4g16780	homeobox-leucine zipper protein/HAT4	*Arabidopsis thaliana*	NC_003075	Chapter 4 Zheng and Nordborg
		Arabidopsis lyrata	Chapter 2/AY174631	Chapter 4
		Capsella rubella	AJ400821	
At4g16800	enoyl-CoA hydratase/ENCOH	*Arabidopsis thaliana*	NC_003075	
		Arabidopsis lyrata	Chapter 2/AF494374	
		Capsella rubella	AJ400821	
At4g21340	Dof zinc finger protein	*Arabidopsis thaliana*	NC_003075	
		Arabidopsis lyrata	E. Kamau, unpublished	
At4g21350	expressed protein	*Arabidopsis thaliana*	NC_003075	
		Arabidopsis lyrata	E. Kamau, unpublished	

At4g21390	expressed protein	*Arabidopsis thaliana*	NC_003075	
		Arabidopsis lyrata	E. Kamau, unpublished	
At4g21430	flavin-containing monooxygenase (FMO) family	*Arabidopsis thaliana*	NC_003075	
		Arabidopsis lyrata	E. Kamau, unpublished	
At4g26090	Disease resistant protein/RPS2	*Arabidopsis thaliana*	NC_003075	
		Arabidopsis lyrata	AF487796	
		Brassica oleracea	AF180355	
At4g28000	hypothetical protein	*Arabidopsis thaliana*	NC_003075	
		Arabidopsis lyrata	AY140458	
		Brassica oleracea	BH951270	
At4g28750	photosystem I subunit PSI-E - like protein	*Arabidopsis thaliana*	NC_003075	
		Arabidopsis lyrata ssp.	this study	
		Brassica oleracea	BH964314, BZ010369	
At4g30950	chloroplast omega-6 fatty acid desaturase	*Arabidopsis thaliana*	AL022198	
		Arabidopsis lyrata ssp.	This study	
		Capsella bursa-pastoris	This study	
		Brassica napus	L29214	
At4g36220	ferulate-5-hydroxylase/FAH1	*Arabidopsis thaliana*	NC_003075	
		Arabidopsis lyrata ssp. petraea	AJ295586	
		Brassica napus	AF214007	
At5g03840	Terminal flower1/TFL1	*Arabidopsis thaliana*	NC_003076	(OLSEN et al. 2002)
		Arabidopsis lyrata	AF466817	
		Brassica napus	AB017525	
At5g04410	NAC2	*Arabidopsis thaliana*	NC_003076	
		Arabidopsis lyrata	AY140486	
		Brassica oleracea	BH922148	
At5g13930	chalcone synthase/CHS	*Arabidopsis thaliana*	NC_003076	
		Arabidopsis lyrata ssp. lyrata	AF112100	
		Arabis glabra	AF112091	

At5G19530[1]	spermine synthase (ACL5)	*Arabidopsis thaliana*	NC_003076	(YOSHIDA et al. 2003)
		Arabidopsis gemmifera	AB076744	
At5g20240	PISTILLATA/PIST	*Arabidopsis thaliana*	NC_003076	(PURUGGANAN and SUDDITH 1999)
		Arabidopsis lyrata ssp.	AF143382	
		Brassica oleracea	BZ520317, BH583322, BH419813, BH650710	
At5g24090	acidic endochitinase/ChiA	*Arabidopsis thaliana*	NC_003076	(KAWABE et al. 1997)
		Arabidopsis lyrata ssp. kawasakiana	AB006072	
		Arabis glabra	AB006071	
At5g42740	phosphoglucose isomerase/PGIC	*Arabidopsis thaliana*	NC_003076	(KAWABE et al. 2000) Zheng and Nordborg Chapter 4
		Arabidopsis lyrata ssp. petraea	AB044969	Chapter 4
		Arabis glabra	AB080908	
At5g61850	floral meristem identity control protein/LFY	*Arabidopsis thaliana*	NC_003076	(OLSEN et al. 2002)
		Arabidopsis lyrata	AF466802	
		Brassica oleracea	BOBOFHA	

1- These genes have not been sequenced in *A. lyrata*, and were only used for polymorphism analysis in *A. thaliana*, using *A. gemmifera* as outgroup.

Appendix A.2 Performance of the maximum likelihood estimator of θ in the presence of intragenic recombination



Appendix A.2 Simulation results studying the behaviour of the maximum likelihood estimator of θ in the presence of intragenic recombination (see Chapter 4). Values shown in black are the rank order maximum likelihood estimates of θ from 1000 coalescent simulations (HUDSON, 2002), and those in grey are the upper- and lower-bounds of the 95% credibility intervals, using the χ^2 approximation . Simulations were run for five loci, using values of sequence length and sample size corresponding to the five nuclear loci studied in Chapter 4 in *A. lyrata ssp. petraea*. Simulations assumed a per base pair value of θ corresponding to the joint estimate for this subspecies (0.0226), shown as an arrow in the figure. Values of the population recombination parameter $\rho=4N_e r$ were the expected values for each locus for ssp. petraea, shown in Table 4.5. Note that most values fall very close to the true value (mean=0.0238, median=0.0237), and that for 1000 simulations, none of the 95% credibility intervals fall outside the true value, making these confidence intervals very conservative in the presence of the expected level of intragenic recombination.

Appendix A.3
Results of coalescent simulations investigating the behaviour of maximum likelihood analysis of multilocus polymorphism and divergence data

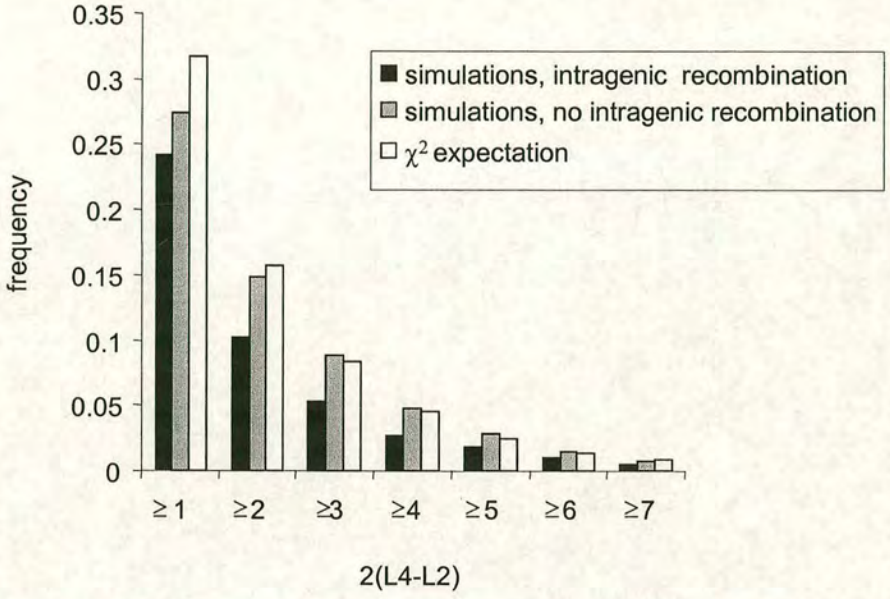


Figure A.3.1 Simulation results studying the behaviour of the likelihood ratio test for selection in the presence and absence of intragenic recombination (see Chapter 6). Results from 1000 neutral coalescent simulations (HUDSON, 2002) were run for 18 loci, matching parameter values to the dataset in Chapter 6. Shown is the frequency distribution for the proportion of simulations showing a difference in log likelihoods calculated for the selection model (model 4 in table 6.3, allowing selection on the AT1G36310) vs. the neutral model (model 2 in table 6.3), assuming free mutation rates across loci. Values of divergence time and theta used in simulations were those estimated using the multilocus HKA procedure for these loci. 1000 coalescent simulations were run for 18 loci to generate simulated datasets for the number of segregating sites. For simulated divergence values, I used coalescent simulations in a sample size of 2 to generate the contribution of ancestral polymorphism, and used simulations of a Poisson process for each locus to get numbers of differences following isolation. For simulations with recombination, I used map-based estimates of recombination and values of theta to generate an expected value for the population parameter $\rho=4Ner$ (see Chapter 4).

Table A.3.1 Simulation results of parameter estimation using the maximum likelihood estimation from polymorphism and divergence (see legend of Figure A.3.1 for simulation details). In general, there appears to be a slight upward bias for divergence time estimation and downward bias for estimation of θ. Inclusion of intragenic recombination gives a less biased estimator of the true values.

Parameter	Value	**no intragenic recombination**		**intragenic recombination**	
		mean	median	mean	median
T	17.83	18.49	18.12	18.25	17.93
θ_1	0.0084	0.0084	0.0083	0.0084	0.0083
θ_2	0.0070	0.0069	0.0067	0.0070	0.0068
θ_3	0.0067	0.0066	0.0065	0.0067	0.0065
θ_4	0.0066	0.0065	0.0063	0.0065	0.0064
θ_5	0.0049	0.0049	0.0047	0.0049	0.0048
θ_6	0.0085	0.0085	0.0083	0.0086	0.0084
θ_7	0.0052	0.0051	0.0050	0.0053	0.0051
θ_8	0.0068	0.0067	0.0066	0.0068	0.0067
θ_9	0.0090	0.0087	0.0085	0.0090	0.0088
θ_{10}	0.0282	0.0279	0.0276	0.0284	0.0279
θ_{11}	0.0105	0.0105	0.0103	0.0105	0.0104
θ_{12}	0.0127	0.0127	0.0124	0.0126	0.0123
θ_{13}	0.0079	0.0079	0.0078	0.0079	0.0077
θ_{14}	0.0077	0.0075	0.0073	0.0076	0.0075
θ_{15}	0.0131	0.0131	0.0129	0.0131	0.0129
θ_{16}	0.0115	0.0113	0.0109	0.0115	0.0114
θ_{17}	0.0087	0.0086	0.0084	0.0088	0.0086
θ_{18}	0.0054	0.0054	0.0053	0.0055	0.0053

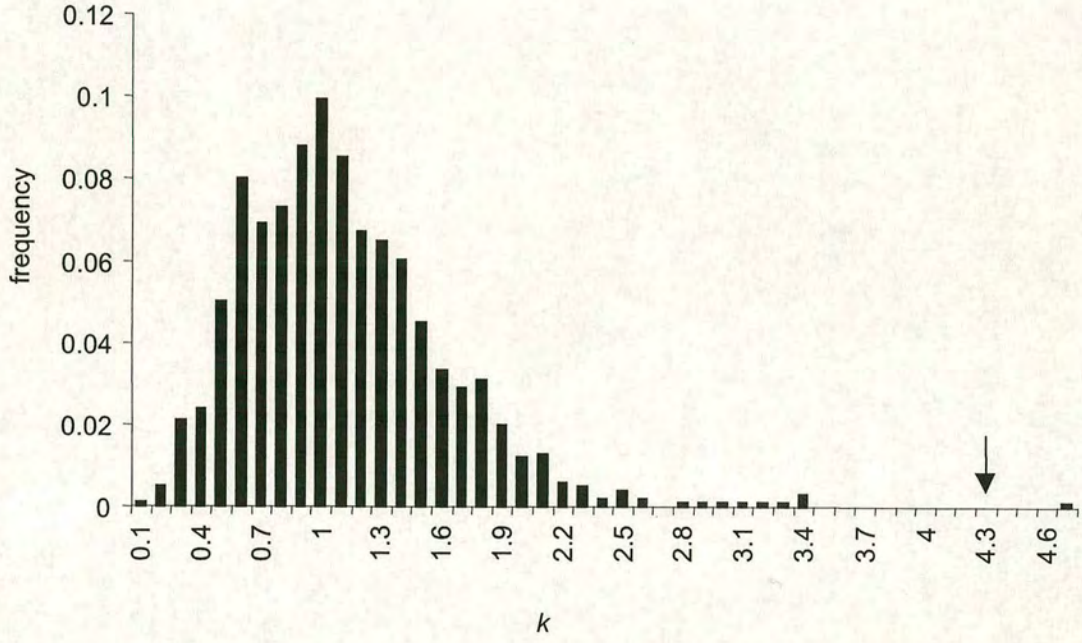


Figure A.3.2 distribution of maximum likelihood estimate of the selection parameter *k*, from 1000 neutral simulations with intragenic recombination (see Chapter 6, Figure A.3.1). Arrow shows value observed for the locus AT1G36310 locus in Chapter 6. Note that, as expected under neutrality, the most frequent values are close to 1 (mean k=1.06, median=0.99).

References

ABBOTT, R. J., AND M.F. GOMES, 1989 Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. Heredity **62:** 411-418.

ACARKAN, A., M. ROSSBERG, M. KOCH and R. SCHMIDT, 2000 Comparative genome analysis reveals extensive conservation of genome organisation for *Arabidopsis thaliana* and *Capsella rubella*. Plant J **23:** 55-62.

AGUADE, M., 2001 Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the FAH1 and F3H genes, in *Arabidopsis thaliana*. Mol Biol Evol **18:** 1-9.

AKASHI, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139:** 1067-1076.

AKASHI, H., 1996 Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in D. melanogaster. Genetics **144:** 1297-1307.

AKASHI, H., 1997 Codon bias evolution in Drosophila. Population genetics of mutation-selection drift. Gene **205:** 269-278.

AKASHI, H., 2001 Gene expression and molecular evolution. Curr Opin Genet Dev **11:** 660-666.

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS and D. J. LIPMAN, 1990 Basic local alignment search tool. J Mol Biol **215:** 403-410.

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389-3402.

ANDOLFATTO, P., 2001 Adaptive hitchhiking effects on genome variability. Curr Opin Genet Dev **11:** 635-641.

ANDOLFATTO, P., and M. PRZEWORSKI, 2000 A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics **156:** 257-268.

ANDOLFATTO, P., and M. PRZEWORSKI, 2001 Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. Genetics **158:** 657-665.

ARABIDOPSIS GENOME INITIATIVE, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796-815.

BARNES, T. M., Y. KOHARA, A. COULSON and S. HEKIMI, 1995 Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. Genetics **141:** 159-179.

BARTOLOME, C., X. MASIDE and B. CHARLESWORTH, 2002 On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. Mol Biol Evol **19:** 926-937.

BARTON, N. H., 2000 Genetic hitchhiking. Philos Trans R Soc Lond B Biol Sci **355:** 1553-1562.

BAUDRY, E., C. KERDELHUE, H. INNAN and W. STEPHAN, 2001 Species and recombination effects on DNA variability in the tomato genus. Genetics **158:** 1725-1735.

BEGUN, D. J., and C. F. AQUADRO, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature **356:** 519-520.

BENNETT, M. D., LEITCH, I.J., HANSON, L., 1998 DNA amounts in two samples of angiosperm weeds. Annals of Botany **Supplement A:** 121-134.

BENSASSON, D., D. A. PETROV, D. X. ZHANG, D. L. HARTL and G. M. HEWITT, 2001 Genomic gigantism: DNA loss is slow in mountain grasshoppers. Mol Biol Evol **18:** 246-253.

BERGE, G., I. NORDAL and G. HESTMARK, 1998 The effect of breeding systems and pollination vectors on the genetic variation of small plant populations within an agricultural landscape. OIKOS **81:** 17-29.

BERGELSON, J., E. STAHL, S. DUDEK and M. KREITMAN, 1998 Genetic variation within and among populations of *Arabidopsis thaliana*. Genetics **148:** 1311-1323.

BETANCOURT, A. J., and D. C. PRESGRAVES, 2002 Linkage limits the power of natural selection in Drosophila. Proc Natl Acad Sci U S A **99:** 13616-13620.

BIEMONT, C., A. TSITRONE, C. VIEIRA and C. HOOGLAND, 1997 Transposable element distribution in Drosophila. Genetics **147:** 1997-1999.

BIRDSELL, J. A., 2002 Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol Biol Evol **19:** 1181-1197.

BOISSINOT, S., A. ENTEZAM and A. V. FURANO, 2001 Selection against deleterious LINE-1-containing loci in the human lineage. Mol Biol Evol **18:** 926-935.

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY and W. STEPHAN, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783-796.

BRENNER, S., M. JOHNSON, J. BRIDGHAM, G. GOLDA, D. H. LLOYD *et al.*, 2000 Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol **18:** 630-634.

BROMHAM, L., D. PENNY, A. RAMBAUT and M. D. HENDY, 2000 The power of relative rates tests depends on the data. J Mol Evol **50:** 296-301.

BULMER, M., 1987 Coevolution of codon usage and transfer RNA abundance. Nature **325:** 728-730.

BUSTAMANTE, C. D., R. NIELSEN, S. A. SAWYER, K. M. OLSEN, M. D. PURUGGANAN *et al.*, 2002 The cost of inbreeding in Arabidopsis. Nature **416:** 531-534.

C. ELEGANS SEQUENCING CONSORTIUM, 1998 Genome sequence of the nematode C. elegans: a platform for investigating biology. Science **282:** 2012-2018.

CARDAZZO, B., P. HAMEL, W. SAKAMOTO, H. WINTZ and G. DUJARDIN, 1998 Isolation of an *Arabidopsis thaliana* cDNA by complementation of a yeast abc1 deletion mutant deficient in complex III respiratory activity. Gene **221:** 117-125.

CARLINI, D. B., Y. CHEN and W. STEPHAN, 2001 The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene

expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr. Genetics **159:** 623-633.

CARLINI, D. B., and W. STEPHAN, 2003 In Vivo Introduction of Unpreferred Synonymous Codons Into the Drosophila Adh Gene Results in Reduced Levels of ADH Protein. Genetics **163:** 239-243.

CARVALHO, A. B., and A. G. CLARK, 1999 Intron size and natural selection. Nature **401:** 344.

CHARLESWORTH, B., 1994 The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet Res **63:** 213-227.

CHARLESWORTH, B., 1996 Background selection and patterns of genetic diversity in Drosophila melanogaster. Genet Res **68:** 131-149.

CHARLESWORTH, B., 1998 Measures of divergence between populations and the effect of forces that reduce variability. Mol Biol Evol **15:** 538-543.

CHARLESWORTH, B., CHARLESWORTH, D., MORGAN, M.T., 1990 Genetic loads and estimates of mutation rates in highly inbred plant populations. Nature **347:** 380-382.

CHARLESWORTH, B., and C. H. LANGLEY, 1989 The population genetics of Drosophila transposable elements. Annu Rev Genet **23:** 251-287.

CHARLESWORTH, B., and C. H. LANGLEY, 1996 The evolution of self-regulated transposition of transposable elements. Genetics **112:** 359-383.

CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289-1303.

CHARLESWORTH, B., M. NORDBORG and D. CHARLESWORTH, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet Res **70:** 155-174.

CHARLESWORTH, D., and B. CHARLESWORTH, 1979 The evolutionary genetics of sexual systems in flowering plants. Proc R Soc Lond B Biol Sci **205:** 513-530.

CHARLESWORTH, D., and B. CHARLESWORTH, 1995 Transposable elements in inbreeding and outbreeding populations. Genetics **140:** 415-417.

CHARLESWORTH, D., B. CHARLESWORTH and M. T. MORGAN, 1995 The pattern of neutral molecular variation under the background selection model. Genetics **141:** 1619-1632.

CHARLESWORTH, D., M.T. MORGAN, AND B. CHARLESWORTH, 1993 Mutation accumulation in finite outbreeding and inbreeding populations. Genet Res **61:** 39-56.

CHARLESWORTH, D., and S. I. WRIGHT, 2001 Breeding systems and genome evolution. Curr Opin Genet Dev **11:** 685-690.

CHIAPELLO, H., F. LISACEK, M. CABOCHE and A. HENAUT, 1998 Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. Gene **209:** GC1-GC38.

COMERON, J. M., 1995 A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J Mol Evol **41:** 1152-1159.

COMERON, J. M., 1999 K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. Bioinformatics **15:** 763-764.

COMERON, J. M., and M. KREITMAN, 1998 The correlation between synonymous and nonsynonymous substitutions in Drosophila: mutation, selection or relaxed constraints? Genetics **150:** 767-775.

COMERON, J. M., and M. KREITMAN, 2000 The correlation between intron length and recombination in Drosophila. Dynamic equilibrium between mutational and selective forces. Genetics **156:** 1175-1190.

COMES, H. P., and J. W. KADEREIT, 1998 The effect of Quaternary climatic changes on plant distribution and evolution. Trends Plant Sci. **3:** 432-438.

COPENHAVER, G. P., K. NICKEL, T. KUROMORI, M. I. BENITO, S. KAUL *et al.*, 1999 Genetic definition and sequence analysis of Arabidopsis centromeres. Science **286:** 2468-2474.

CUTTER, A. D., and B. A. PAYSEUR, 2003 Selection at Linked Sites in the Partial Selfer *Caenorhabditis elegans*. Mol Biol Evol **20:** 665-673.

DABORN, P. J., J. L. YEN, M. R. BOGWITZ, G. LE GOFF, E. FEIL *et al.*, 2002 A single p450 allele associated with insecticide resistance in Drosophila. Science **297:** 2253-2256.

DESALLE, R. A. T., A.R., 1988 Founder effects and the rate of mitochondrial DNA evolution in Hawaiian Drosophila. Evolution **42:** 1076-1084.

DEVOS, K. M., J. K. BROWN and J. L. BENNETZEN, 2002 Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res **12:** 1075-1079.

DUNN, K. A., J. P. BIELAWSKI and Z. YANG, 2001 Substitution rates in Drosophila nuclear genes: implications for translational selection. Genetics **157:** 295-305.

DURET, L., 2000 tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. Trends Genet **16:** 287-289.

DURET, L., and N. GALTIER, 2000 The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. Mol Biol Evol **17:** 1620-1625.

DURET, L., G. MARAIS and C. BIEMONT, 2000 Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*. Genetics **156:** 1661-1669.

DURET, L., and D. MOUCHIROUD, 1999 Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci U S A **96:** 4482-4487.

DURET, L., and D. MOUCHIROUD, 2000 Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol **17:** 68-74.

EICKBUSH, T. H., and A. V. FURANO, 2002 Fruit flies and humans respond differently to retrotransposons. Curr Opin Genet Dev **12:** 669-674.

EYRE-WALKER, A., 1999 Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics **152:** 675-683.

FAY, J. C., and C. I. WU, 1999 A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Mol Biol Evol **16:** 1003-1005.

FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405-1413.

FAY, J. C., G. J. WYCKOFF and C. I. WU, 2002 Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature **415:** 1024-1026.

FILATOV, D. A., and D. CHARLESWORTH, 1999 DNA polymorphism, haplotype structure and balancing selection in the Leavenworthia PgiC locus. Genetics **153:** 1423-1434.

FRISSE, L., R. R. HUDSON, A. BARTOSZEWICZ, J. D. WALL, J. DONFACK *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet **69:** 831-843.

FU, H., Z. ZHENG and H. K. DOONER, 2002 Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc Natl Acad Sci U S A **99:** 1082-1087.

FULLERTON, S. M., A. BERNARDO CARVALHO and A. G. CLARK, 2001 Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol **18:** 1139-1142.

FUNK, D. J., J. J. WERNEGREEN and N. A. MORAN, 2001 Intraspecific variation in symbiont genomes: bottlenecks and the aphid- buchnera association. Genetics **157:** 477-489.

GALTIER, N., G. PIGANEAU, D. MOUCHIROUD and L. DURET, 2001 GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics **159:** 907-911.

GAUT, B. S., B. R. MORTON, B. C. MCCAIG and M. T. CLEGG, 1996 Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc Natl Acad Sci U S A **93:** 10274-10279.

GENDREL, A. V., Z. LIPPMAN, C. YORDAN, V. COLOT and R. A. MARTIENSSEN, 2002 Dependence of heterochromatic histone H3 methylation patterns on the Arabidopsis gene DDM1. Science **297:** 1871-1873.

GRAUSTEIN, A., J. M. GASPAR, J. R. WALTERS and M. F. PALOPOLI, 2002 Levels of DNA polymorphism vary with mating system in the nematode genus caenorhabditis. Genetics **161:** 99-107.

HAGENBLAD, J., and M. NORDBORG, 2002 Sequence Variation and Haplotype Structure Surrounding the Flowering Time Locus FRI in *Arabidopsis thaliana*. Genetics **161:** 289-298.

HARRISON, P. M., N. ECHOLS and M. B. GERSTEIN, 2001 Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. Nucleic Acids Res **29:** 818-830.

HAUBOLD, B., J. KROYMANN, A. RATZKA, T. MITCHELL-OLDS and T. WIEHE, 2002 Recombination and Gene Conversion in a 170-kb Genomic Region of *Arabidopsis thaliana*. Genetics **161:** 1269-1278.

HAUPT, W., T. C. FISCHER, S. WINDERL, P. FRANSZ and R. A. TORRES-RUIZ, 2001 The CENTROMERE1 (CEN1) region of *Arabidopsis thaliana*: architecture and functional impact of chromatin. Plant J **27:** 285-296.

HAUSER, M. T., B. HARR and C. SCHLOTTERER, 2001 Trichome Distribution in *Arabidopsis thaliana* and its Close Relative *Arabidopsis lyrata*: Molecular Analysis of the Candidate Gene GLABROUS1. Mol Biol Evol **18:** 1754-1763.

HELLER, J., and J. MAYNARD SMITH, 1979 Does Muller's ratchet work with selfing? Genetical Research **32:** 289-294.

HELLMANN, I., I. EBERSBERGER, S. E. PTAK, S. PAABO and M. PRZEWORSKI, 2003 A neutral explanation for the correlation of diversity with recombination rates in humans. Am J Hum Genet **72:** 1527-1535.

HILL, W. G., and A. ROBERTSON, 1966 The effect of linkage on limits to artificial selection. Genet Res **8:** 269-294.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process., pp. 1-44 in *Oxford Surveys in evolutionary biology*, edited by J. A. D. FUTUYMA. Oxford University Press, New York.

HUDSON, R. R., 2000 A new statistic for detecting genetic differentiation. Genetics **155:** 2011-2014.

HUDSON, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805-1817.

HUDSON, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18:** 337-338.

HUDSON, R. R., D. D. BOOS and N. L. KAPLAN, 1992a A statistical test for detecting geographic subdivision. Mol Biol Evol **9:** 138-151.

HUDSON, R. R., and N. L. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147-164.

HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153-159.

HUDSON, R. R., M. SLATKIN and W. P. MADDISON, 1992b Estimation of levels of gene flow from DNA sequence data. Genetics **132:** 583-589.

IKEMURA, T., 1985 Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol **2:** 13-34.

INGRAM, G. C., S. DOYLE, R. CARPENTER, E. A. SCHULTZ, R. SIMON *et al.*, 1997 Dual role for fimbriata in regulating floral homeotic genes and cell division in Antirrhinum. Embo J **16:** 6521-6534.

INNAN, H., and W. STEPHAN, 2000 The coalescent in an exponentially growing metapopulation and its application to *Arabidopsis thaliana*. Genetics **155:** 2015-2019.

INNAN, H., F. TAJIMA, R. TERAUCHI and N. T. MIYASHITA, 1996 Intragenic recombination in the Adh locus of the wild plant *Arabidopsis thaliana*. Genetics **143:** 1761-1770.

JAKUBCZAK, J. L., W. D. BURKE and T. H. EICKBUSH, 1991 Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. Proc Natl Acad Sci U S A **88:** 3295-3299.

JOHNSON, K. P., and J. SEGER, 2001 Elevated rates of nonsynonymous substitution in island birds. Mol Biol Evol **18:** 874-881.

JUNGHANS, H., and M. METZLAFF, 1990 A simple and rapid method for the preparation of total plant DNA. Biotechniques **8:** 176.

KABACK, D. B., D. BARBER, J. MAHON, J. LAMB and J. YOU, 1999 Chromosome size-dependent control of meiotic reciprocal recombination in *Saccharomyces cerevisiae*: the role of crossover interference. Genetics **152:** 1475-1486.

KANAYA, S., Y. YAMADA, M. KINOUCHI, Y. KUDO and T. IKEMURA, 2001 Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J Mol Evol **53:** 290-298.

KAPITONOV, V. V., and J. JURKA, 2001 Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci U S A **98:** 8714-8719.

KATTI, M. V., P. K. RANJEKAR and V. S. GUPTA, 2001 Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol **18:** 1161-1167.

KAWABE, A., H. INNAN, R. TERAUCHI and N. T. MIYASHITA, 1997 Nucleotide polymorphism in the acidic chitinase locus (ChiA) region of the wild plant *Arabidopsis thaliana*. Mol Biol Evol **14:** 1303-1315.

KAWABE, A., and N. T. MIYASHITA, 1999a DNA variation in the basic chitinase locus (ChiB) region of the wild plant *Arabidopsis thaliana*. Genetics **153:** 1445-1453.

KAWABE, A., and N. T. MIYASHITA, 1999b DNA variation in the basic chitinase locus (ChiB) region of the wild plant *Arabidopsis thaliana*. Genetics **153:** 1445-1453.

KAWABE, A., K. YAMANE and N. T. MIYASHITA, 2000 DNA polymorphism at the cytosolic phosphoglucose isomerase (PgiC) locus of the wild plant *Arabidopsis thaliana*. Genetics **156:** 1339-1347.

KEIGHTLEY, P. D., and A. EYRE-WALKER, 2000 Deleterious mutations and the evolution of sex. Science **290:** 331-333.

KIM, Y., and W. STEPHAN, 2000 Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics **155:** 1415-1427.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

KIRIK, A., S. SALOMON and H. PUCHTA, 2000 Species-specific double-strand break repair and genome evolution in plants. Embo J **19:** 5562-5566.

KLIEBENSTEIN, D. J., J. KROYMANN, P. BROWN, A. FIGUTH, D. PEDERSEN *et al.*, 2001 Genetic control of natural variation in Arabidopsis glucosinolate accumulation. Plant Physiol **126:** 811-825.

KLIMAN, R. M., P. ANDOLFATTO, J. A. COYNE, F. DEPAULIS, M. KREITMAN *et al.*, 2000 The population genetics of the origin and divergence of the *Drosophila simulans* complex species. Genetics **156:** 1913-1931.

KLIMAN, R. M., and J. HEY, 1993 Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol Biol Evol **10:** 1239-1258.

KLIMAN, R. M., and J. HEY, 2003 Hill-Robertson interference in *Drosophila melanogaster*: reply to Marais, Mouchiroud and Duret. Genet Res **81:** 89-90.

KOCH, M., B. HAUBOLD and T. MITCHELL-OLDS, 2001 Molecular systematics of the Brassicaceae: evidence from coding plastidic matK and nuclear Chs sequences. Am J Bot **88:** 534-544.

KOCH, M. A., B. HAUBOLD and T. MITCHELL-OLDS, 2000 Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). Mol Biol Evol **17:** 1483-1498.

KUITTINEN, H., and M. AGUADE, 2000 Nucleotide variation at the CHALCONE ISOMERASE locus in *Arabidopsis thaliana*. Genetics **155:** 863-872.

KUITTINEN, H., A. DE HAAN, C. VOGL, J. LEPPALA, T. MITCHELL-OLDS *et al.*, *unpublished* A genetic map of *Arabdiopsis lyrata.*

KUMEKAWA, N., T. HOSOUCHI, H. TSURUOKA and H. KOTANI, 2000 The size and sequence organization of the centromeric region of *Arabidopsis thaliana* chromosome 5. DNA Res **7:** 315-321.

KUSABA, M., K. DWYER, J. HENDERSHOT, J. VREBALOV, J. B. NASRALLAH *et al.*, 2001 Self-incompatibility in the genus Arabidopsis: characterization of the S locus in the outcrossing A. lyrata and its autogamous relative *A. thaliana*. Plant Cell **13:** 627-643.

LANDER, E. S., L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860-921.

LANGLEY, C. H., B. P. LAZZARO, W. PHILLIPS, E. HEIKKINEN and J. M. BRAVERMAN, 2000 Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the *Drosophila melanogaster* X chromosome. Genetics **156:** 1837-1852.

LANGLEY, C. H., J. MACDONALD, N. MIYASHITA and M. AGUADE, 1993 Lack of correlation between interspecific divergence and intraspecific polymorphism at the suppressor of forked region in *Drosophila melanogaster* and *Drosophila simulans*. Proc Natl Acad Sci U S A **90:** 1800-1803.

LANGLEY, C. H., E. MONTGOMERY, R. HUDSON, N. KAPLAN and B. CHARLESWORTH, 1988 On the role of unequal exchange in the containment of transposable element copy number. Genet Res **52:** 223-235.

LE, Q. H., S. WRIGHT, Z. YU and T. BUREAU, 2000 Transposon diversity in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A **97:** 7376-7381.

LERCHER, M. J., and L. D. HURST, 2002 Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet **18:** 337-340.

LERCHER, M. J., N. G. SMITH, A. EYRE-WALKER and L. D. HURST, 2002 The evolution of isochores: evidence from SNP frequency distributions. Genetics **162:** 1805-1810.

LICHTEN, M., and A. S. GOLDMAN, 1995 Meiotic recombination hotspots. Annu Rev Genet **29:** 423-444.

LIU, F., D. CHARLESWORTH and M. KREITMAN, 1999 The effect of mating system differences on nucleotide diversity at the phosphoglucose isomerase locus in the plant genus Leavenworthia. Genetics **151:** 343-357.

LIU, F., L. ZHANG and D. CHARLESWORTH, 1998 Genetic diversity in Leavenworthia populations with different inbreeding levels. Proc R Soc Lond B Biol Sci **265:** 293-301.

LONG, A. D., R. F. LYMAN, A. H. MORGAN, C. H. LANGLEY and T. F. MACKAY, 2000 Both naturally occurring insertions of transposable elements and intermediate frequency polymorphisms at the achaete-scute complex are associated with variation in bristle number in *Drosophila melanogaster*. Genetics **154:** 1255-1269.

LYNCH, M., and J. L. BLANCHARD, 1998 Deleterious mutation accumulation in organelle genomes. Genetica **103:** 29-39.

MALIK, H. S., and S. HENIKOFF, 2002 Conflict begets complexity: the evolution of centromeres. Curr Opin Genet Dev **12:** 711-718.

MARAIS, G., 2003 Biased gene conversion: implications for genome and sex evolution. Trends Genet **19:** 330-338.

MARAIS, G., D. MOUCHIROUD and L. DURET, 2001 Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci U S A **98:** 5688-5692.

MARAIS, G., and G. PIGANEAU, 2002 Hill-Robertson interference is a minor determinant of variations in codon bias across *Drosophila melanogaster* and *Caenorhabditis elegans* genomes. Mol Biol Evol **19:** 1399-1406.

MATZKE, M. A., O. M. SCHEID and A. J. MATZKE, 1999 Rapid structural and epigenetic changes in polyploid and aneuploid genomes. Bioessays **21:** 761-767.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the Adh locus in Drosophila. Nature **351:** 652-654.

MCVEAN, G. A., and B. CHARLESWORTH, 2000 The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155:** 929-944.

MCVEAN, G. A., and J. VIEIRA, 2001 Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics **157:** 245-257.

MEDSTRAND, P., L. N. VAN DE LAGEMAAT and D. L. MAGER, 2002 Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res **12:** 1483-1495.

MORGAN, M. T., 2001 Transposable element number in mixed mating populations. Genet Res **77:** 261-275.

MORIYAMA, E. N., and J. R. POWELL, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol Biol Evol **13:** 261-277.

MORIYAMA, E. N., and J. R. POWELL, 1997 Codon usage bias and tRNA abundance in Drosophila. J Mol Evol **45:** 514-523.

NACHMAN, M. W., 1997 Patterns of DNA variability at X-linked loci in *Mus domesticus*. Genetics **147:** 1303-1316.

NACHMAN, M. W., 2001 Single nucleotide polymorphisms and recombination rate in humans. Trends Genet **17:** 481-485.

NACHMAN, M. W., V. L. BAUER, S. L. CROWELL and C. F. AQUADRO, 1998 DNA variability and recombination rates at X-linked loci in humans. Genetics **150:** 1133-1141.

NEKRUTENKO, A. A. L., W.-H., 2001 Transposable elements are found in a large number of human protein-coding genes. Trends Genet **17:** 619-621.

NICHOLAS, K. B., NICHOLAS, H.B. JR., DEERFIELD, D.W. II., 1997 GeneDoc: Analysis and Visualization of Genetic Variation. EMBNEW.NEWS **4:** 14.

NORDBORG, M., 2000a Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics **154:** 923-929.

NORDBORG, M., 2000b On Detecting Ancient Admixture in *Genes, Fossils, and Behaviour: An Integrated Approach to Human Evolution.*, edited by P. DONNELLY. IOS Press, Amsterdam.

NORDBORG, M., J. O. BOREVITZ, J. BERGELSON, C. C. BERRY, J. CHORY *et al.*, 2002a The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet **7:** 7.

NUZHDIN, S. V., E. G. PASYUKOVA and T. F. MACKAY, 1996 Positive association between copia transposition rate and copy number in *Drosophila melanogaster*. Proc R Soc Lond B Biol Sci **263:** 823-831.

NUZHDIN, S. V., E. G. PASYUKOVA and T. F. MACKAY, 1997 Accumulation of transposable elements in laboratory lines of *Drosophila melanogaster*. Genetica **100:** 167-175.

OHTA, T., 1992 The nearly neutral model of molecular evolution. Annual Review of Ecology and Systematics **23:** 263-286.

OHTA, T., 1993a Amino acid substitution at the Adh locus of Drosophila is facilitated by small population size. Proc Natl Acad Sci U S A **90:** 4548-4551.

OHTA, T., 1993b An examination of the generation-time effect on molecular evolution. Proc Natl Acad Sci U S A **90:** 10676-10680.

OLSEN, K. M., A. WOMACK, A. R. GARRETT, J. I. SUDDITH and M. D. PURUGGANAN, 2002 Contrasting Evolutionary Forces in the *Arabidopsis thaliana* Floral Developmental Pathway. Genetics **160:** 1641-1650.

PANNELL, J. R., and B. CHARLESWORTH, 2000 Effects of metapopulation processes on measures of genetic diversity. Philos Trans R Soc Lond B Biol Sci **355:** 1851-1864.

PAVLICEK, A., O. CLAY and G. BERNARDI, 2002 Transposable elements encoding functional proteins: pitfalls in unprocessed genomic data? FEBS Lett **523:** 252-253.

PELISSIER, T., S. TUTOIS, J. M. DERAGON, S. TOURMENTE, S. GENESTIER *et al.*, 1995 Athila, a new retroelement from *Arabidopsis thaliana*. Plant Mol Biol **29:** 441-452.

PERCUDANI, R., 2001 Restricted wobble rules for eukaryotic genomes. Trends Genet **17:** 133-135.

PETROV, D. A., and D. L. HARTL, 1997 Trash DNA is what gets thrown away: high rate of DNA loss in Drosophila. Gene **205:** 279-289.

PETROV, D. A., E. R. LOZOVSKAYA and D. L. HARTL, 1996 High intrinsic rate of DNA loss in Drosophila. Nature **384:** 346-349.

POLLAK, E., 1987 On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics **117:** 353-360.

PRITCHARD, J. K., and M. PRZEWORSKI, 2001 Linkage disequilibrium in humans: models and data. Am J Hum Genet **69:** 1-14.

PRZEWORSKI, M., and J. D. WALL, 2001 Why is there so little intragenic linkage disequilibrium in humans? Genet Res **77:** 143-151.

PURUGGANAN, M. D., and J. I. SUDDITH, 1998 Molecular population genetics of the Arabidopsis CAULIFLOWER regulatory gene: nonneutral evolution and naturally occurring variation in floral homeotic function. Proc Natl Acad Sci U S A **95:** 8130-8134.

PURUGGANAN, M. D., and J. I. SUDDITH, 1999 Molecular population genetics of floral homeotic loci. Departures from the equilibrium-neutral model at the APETALA3 and PISTILLATA genes of *Arabidopsis thaliana*. Genetics **151:** 839-848.

RODRIGUEZ-TRELLES, F., R. TARRIO and F. J. AYALA, 1999 Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. Genetics **153:** 339-350.

ROZAS, J., and R. ROZAS, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174-175.

SANMIGUEL, P., B. S. GAUT, A. TIKHONOV, Y. NAKAJIMA and J. L. BENNETZEN, 1998 The paleontology of intergene retrotransposons of maize. Nat Genet **20:** 43-45.

SAVOLAINEN, O., C. H. LANGLEY, B. P. LAZZARO and H. FR, 2000 Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. Mol Biol Evol **17:** 645-655.

SAWYER, S. A., and D. L. HARTL, 1992 Population genetics of polymorphism and divergence. Genetics **132:** 1161-1176.

SCHIERUP, M. H., B. K. MABLE, P. AWADALLA and D. CHARLESWORTH, 2001 Identification and characterization of a polymorphic receptor kinase gene linked to the self-incompatibility locus of *Arabidopsis lyrata*. Genetics **158:** 387-399.

SCHNABLE, P. S., A. P. HSIA and B. J. NIKOLAU, 1998 Genetic recombination in plants. Curr Opin Plant Biol **1:** 123-129.

SCHULTZ, S. T., M. LYNCH and J. H. WILLIS, 1999 Spontaneous deleterious mutation in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A **96:** 11393-11398.

SHARBEL, T. F., B. HAUBOLD and T. MITCHELL-OLDS, 2000 Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. Mol Ecol **9:** 2109-2118.

SINGER, T., C. YORDAN and R. A. MARTIENSSEN, 2001 Robertson's Mutator transposons in *A. thaliana* are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). Genes Dev **15:** 591-602.

SLATKIN, M., 1994 Linkage disequilibrium in growing and stable populations. Genetics **137:** 331-336.

SMITH, N. G., and A. EYRE-WALKER, 2001 Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. Mol Biol Evol **18:** 982-986.

SMITH, N. G., and A. EYRE-WALKER, 2002 Adaptive protein evolution in Drosophila. Nature **415:** 1022-1024.

STAHL, E. A., G. DWYER, R. MAURICIO, M. KREITMAN and J. BERGELSON, 1999 Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature **400:** 667-671.

STENICO, M., A. T. LLOYD and P. M. SHARP, 1994 Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucleic Acids Res **22:** 2437-2446.

STEPHAN, W., and C. H. LANGLEY, 1998 DNA polymorphism in Lycopersicon and crossing-over per physical length. Genetics **150:** 1585-1593.

STROBECK, C., 1987 Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics **117:** 149-153.

SURZYCKI, S. A., and W. R. BELKNAP, 1999 Characterization of repetitive DNA elements in Arabidopsis. J Mol Evol **48:** 684-691.

TACHIDA, H., 2000 Molecular evolution in a multisite nearly neutral mutation model. J Mol Evol **50:** 69-81.

TAJIMA, F., 1989a The effect of change in population size on DNA polymorphism. Genetics **123:** 597-601.

TAJIMA, F., 1989b Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595.

TAJIMA, F., 1993 Simple methods for testing the molecular evolutionary clock hypothesis. Genetics **135:** 599-607.

TAKAHATA, N., Y. SATTA and J. KLEIN, 1995 Divergence time and population size in the lineage leading to modern humans. Theor Popul Biol **48:** 198-221.

TAKANO-SHIMIZU, T., 1999 Local recombination and mutation effects on molecular evolution in Drosophila. Genetics **153:** 1285-1296.

TAKANO-SHIMIZU, T., 2001 Local changes in GC/AT substitution biases and in crossover frequencies on Drosophila chromosomes. Mol Biol Evol **18:** 606-619.

TAKEBAYASHI, N., and P. L. MORRELL, 2001 Is self-fertilization an evolutionary dead end? Revisiting an old hypothesis with genetic theories and a macroevolutionary approach. Am J Bot **88:** 1143-1150.

TARRIO, R., F. RODRIGUEZ-TRELLES and F. J. AYALA, 2001 Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. Mol Biol Evol **18:** 1464-1473.

TENAILLON, M. I., M. C. SAWKINS, L. K. ANDERSON, S. M. STACK, J. DOEBLEY *et al.*, 2002 Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays ssp. mays L.*). Genetics **162:** 1401-1413.

TENAILLON, M. I., M. C. SAWKINS, A. D. LONG, R. L. GAUT, J. F. DOEBLEY *et al.*, 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays ssp. mays L.*). Proc Natl Acad Sci U S A **24:** 24.

THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22:** 4673-4680.

TIFFIN, P., and M. W. HAHN, 2002 Coding sequence divergence between two closely related plant species: *Arabidopsis thaliana* and *Brassica rapa ssp. pekinensis*. J Mol Evol **54:** 746-753.

TODOKORO, S., R. TERAUCHI and S. KAWANO, 1995 Microsatellite polymorphisms in natural populations of *Arabidopsis thaliana* in Japan. Japanese J Genetics **70:** 543-554.

VIRGIN, J. B., and J. P. BAILEY, 1998 The M26 hotspot of *Schizosaccharomyces pombe* stimulates meiotic ectopic recombination and chromosomal rearrangements. Genetics **149:** 1191-1204.

WAKELEY, J., 2003 Polymorphism and divergence for island-model species. Genetics **163:** 411-420.

WAKELEY, J., and N. ALIACAR, 2001 Gene genealogies in a metapopulation. Genetics **159:** 893-905.

WALL, J., 1999 Recombination and the power of statistical tests of neutrality. Genet Res **74:** 65-79.

WALL, J. D., 2000 A comparison of estimators of the population recombination rate. Mol Biol Evol **17:** 156-163.

WALL, J. D., 2003 Estimating ancestral population sizes and divergence times. Genetics **163:** 395-404.

WALL, J. D., P. ANDOLFATTO and M. PRZEWORSKI, 2002 Testing models of selection and demography in *Drosophila simulans*. Genetics **162:** 203-216.

WALTER, R., and B. K. EPPERSON, 2001 Geographic pattern of genetic variation in *Pinus resinosa*: area of greatest diversity is not the origin of postglacial populations. Mol Ecol **10:** 103-111.

WANG, X., X. SHI and B. HAO, 2002 The transfer RNA genes in *Oryza sativa* L. ssp. *indica*. Science in China (Series C) **45:** 504-511.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor Popul Biol **7:** 256-276.

WEINREICH, D. M., and D. M. RAND, 2000 Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondrial genomes. Genetics **156:** 385-399.

WELCH, D. B., and M. S. MESELSON, 2001 Rates of nucleotide substitution in sexual and anciently asexual rotifers. Proc Natl Acad Sci U S A **98:** 6720-6724.

WERNEGREEN, J. J., and N. A. MORAN, 1999 Evidence for genetic drift in endosymbionts (Buchnera): analyses of protein-coding genes. Mol Biol Evol **16:** 83-97.

WHITLOCK, M. C., and N. H. BARTON, 1997 The effective size of a subdivided population. Genetics **146:** 427-441.

WHITLOCK, M. C., and D. E. MCCAULEY, 1999 Indirect measures of gene flow and migration: FST not equal to 1/(4Nm + 1). Heredity **82:** 117-125.

WILSON, R. K., 1999 How the worm was won. The *C. elegans* genome sequencing project. Trends Genet **15:** 51-58.

WOLFE, K. H., W. H. LI and P. M. SHARP, 1987 Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A **84:** 9054-9058.

WRIGHT, S., and D. FINNEGAN, 2001 Genome evolution: sex and the transposable element. Curr Biol **11:** R296-299.

WRIGHT, S. I., Q. H. LE, D. J. SCHOEN and T. E. BUREAU, 2001 Population Dynamics of an Ac-like Transposable Element in Self- and Cross-Pollinating Arabidopsis. Genetics **158:** 1279-1288.

WRIGHT, S. I., and D. J. SCHOEN, 1999 Transposon dynamics and the breeding system. Genetica **107:** 139-148.

YANG, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:** 555-556.

YOSHIDA, K., T. KAMIYA, A. KAWABE and N. T. MIYASHITA, 2003 DNA polymorphism at the ACAULIS5 locus of the wild plant Arabidopsis thaliana. Genes Genet Syst **78:** 11-21.

YOUNG, N. D., and C. W. DEPAMPHILIS, 2000 Purifying selection detected in the plastid gene matK and flanking ribozyme regions within a group II intron of nonphotosynthetic plants. Mol Biol Evol **17:** 1933-1941.

YU, Z., S. I. WRIGHT and T. E. BUREAU, 2000 Mutator-like elements in *Arabidopsis thaliana*. Structure, diversity and evolution. Genetics **156:** 2019-2031.